COMPUTERIZED SCREENING SYSTEM: AN APPLICATION OF THE RELEVANT COMPARISON TEST IN THE DETECTION OF DECEPTION

by

Jessica Dawn Jewell

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Educational Psychology

The University of Utah

May 2016

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Jessica Dawn Jewell**

has been approved by the following supervisory committee members:

John C. Kircher	, Chair	**12/10/2015** Date Approved
Amy Jo Metz	, Member	**12/10/2015** Date Approved
Jason Burrow-Sanchez	, Member	**12/10/2015** Date Approved
Michael K. Gardner	, Member	**12/10/2015** Date Approved
Lauren Weitzman	, Member	**12/10/2015** Date Approved

and by **Anne Cook** , Chair/Dean of

the Department/College/School of **Educational Psychology**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

The current study evaluated the Computerized Screening System (CSS) for ports of entry. Additional primary objectives included evaluating the Relevant Comparison Test (RCT) for use at ports of entry, less invasive alternatives to skin conductance and the cardiograph, and alternative statistical methods for classification.

Data were collected in two phases. Complete sets of recordings were obtained from 169 Phase 1 participants and 185 Phase 2 participants (*N* = 354). Participants were either guilty (*n* = 230) or innocent (*n* = 124) of committing a mock crime. Guilty participants transported a substance that appeared to be illegal drugs (*n* = 119), or they transported a device that appeared to be a bomb (*n* = 111). When the participant reported to the laboratory, a research assistant initiated a computer program that presented prerecorded auditory instructions and test questions to the participant.

The computer administered a test entitled the Relevant Comparison Test (RCT) that was developed specifically for this project. The RCT included 12 relevant questions about the bomb condition, 12 relevant questions about the drug condition, and 24 neutral questions. Respiration, electrodermal, cardiovascular, and pupil reactions were recorded continuously throughout the test.

As expected, guilty participants who transported the drugs reacted more strongly to questions about the drugs than to questions about the bomb. Participants who were guilty of transporting the bomb reacted more strongly to questions about the bomb. Innocent participants reacted similarly to questions about the drugs and the bomb, although there was a tendency for some innocent participants to react to questions about the drugs. Increases in

diastolic blood pressure and systolic blood pressure were most diagnostic of group membership behind skin conductance. Contrary to expectations, pupil measures did not perform as well as skin conductance measures, and traditional discriminant analysis was more effective than the computer-intensive bagging and boosting classification techniques.

TABLE OF CONTENTS

ABSTRACT iii

LIST OF FIGURES vii

LIST OF TABLES viii

Chapters

1 INTRODUCTION 1

Decision Algorithms 6
Research Objectives 8

2 METHOD 9

Participants 9
Procedures 10
Relevant Comparison Test (RCT) 12
Apparatus 13
Summary of Differences between Phase 1 and Phase 2 15
Computer Measurements of Physiological Waveforms 15
Feature Extraction 18
Within-Subject Standardization of Physiological Features 20
Indices of Differential Reactivity 20
Classification Techniques 21

3 RESULTS 22

Preliminary Analysis: Effects of Procedural Changes between Phase 1 and Phase 2 22
Analysis of Physiological Activity 24
Bivariate Analyses 36
Dependent Variables 36
Independent Variables 37
Classification Analyses 40
Standardization and Validation Samples 41
Validation Sample 42
Ensemble Classification 42
Effects of Trimming Branches from Boosting Decision Trees 46

4 DISCUSSION 47

Relevant Comparison Test 47
Automation 49
Less Invasive Alternatives to Skin Conductance and the Cardiograph 50
Pupil Size 50
Cardiovascular Measures 51
Ensemble Classification Methods 52
Limitations 54
General Conclusions 56

Appendices

A COMPUTERIZED SCREENING SYSTEM INCLUSION /EXCLUSION CRITERIA 57

B AUTOMATED POLYGRAPH SCRIPT 58

REFERENCES 59

LIST OF FIGURES

Figure

1. Mean second-by-second respiration line length for innocent, drug, and bomb groups 27

2. Mean second-by-second change in standardized skin conductance for innocent, drug, and bomb groups 28

3. Mean second-by-second change in finger pulse amplitude for innocent, drug, and bomb groups 29

4. Mean second-by-second systolic blood pressure for innocent, drug, and bomb groups 30

5. Mean second-by-second diastolic blood pressure for innocent, drug, and bomb groups 31

6. Mean second-by-second change in total peripheral resistance (TPR) for innocent, drug, and bomb groups 32

7. Mean second-by-second heart rate for innocent, drug, and bomb groups 33

8. Mean second-by-second change in pupil diameter for innocent, drug, and bomb groups 34

LIST OF TABLES

Table

1. Procedural differences between Phase 1 and Phase 2 15

2. Effects of changes from Phase 1 to Phase 2 on physiological features: Condition X Question Type X Phase interaction effects 23

3. Repeated measures analysis of variance results for Phase 1 and Phase 2 combined 26

4. Correlations between physiological features and group membership indicator variables (N=354) 38

5. Standardized discriminant function coefficients for functions 1 and 2 in the ensemble standardization sample 42

6. Classifications of participants in the standardization sample 42

7. Classifications of participants in the validation sample 43

8. Percent of cases classified correctly for discriminant function analysis (DFA) and ensemble methods 44

CHAPTER 1

INTRODUCTION

Concerns regarding our national security have been in the forefront of immigration policy since the 9/11 attacks. Over 350 million people entered the United States in 2013, including 102 million air passengers and crew and 242 million land travelers. About 205,000 travelers were denied admission to the United States, and 24,000 were arrested on criminal warrants. An unknown number of travelers entered the United States through fraudulent means. Immigration enforcement officers remain the primary line of defense against transnational crime, such as drug trafficking, as well as international terrorism (Seghetti, 2014). Improvements to inspection processes at ports of entry were one of three major recommendations made by the 9/11 Commission in their comprehensive report to address security lapses that allowed the 19 Al Qaeda operatives involved in the attack to enter the country legally (National Commission on Terrorist Attacks upon the United States, 2004). Their report encouraged the U.S. Government to employ an automated electronic screening system that includes biometric identifiers, such as fingerprints and digital photographs to confirm identity and legality of entrance to the United States.

Improvement in pretravel screening measures could decrease unlawful attempts to enter the United States, including visa applications and interviews, as could improvements to primary inspections, such as verifying travel documents. Questions or concerns at the primary inspection level result in a secondary inspection, a more thorough review of documentation, physical search, and/or interview by Homeland Security personnel. These personnel were

responsible for interviewing over 5 million travelers flagged for secondary inspection in 2013 (Seghetti, 2014). While evidence exists to support a claim that trained personnel can detect lies better than the average person (O'Sullivan et al., 2009), the bulk of the literature concludes that most people, including police officers, can detect lies with approximately 55% accuracy, which is insufficient to assure the security of our nation (Aamodt & Custer, 2006).

Polygraph examinations are a viable alternative to in-person interviews designed to detect intent of wrongdoing. Polygraph tests have wide-ranging and useful applications in employment screening, criminal investigations, and periodic testing of personnel with security clearances (Krapohl, 2001). However, the process for training polygraph examiners is lengthy and expensive, and the validity of tests partly depends on the expertise of the polygraph examiner who constructs test questions, completes a pretest interview, administers the test, and interprets the results. Automating the analysis of polygraph recordings has been shown to reduce decision error (Dollins, Krapohl, & Dutton, 2000; Kircher & Raskin, 2001). Automating the analysis of the physiological data removes variance due to polygraph examiners' scoring of the polygraph, but it does not control for individual differences among examiners in age, sex, ethnicity, training, experience, or interview skills, or any interactions of those factors with individual differences among examinees.

Automated polygraph applications could completely eliminate the need for costly training of personnel; a technician could administer an automated polygraph examination with very little training. In addition, automated polygraph systems could draw questions from a large bank of appropriate questions, thereby increasing the generalizability of the polygraph to a variety of circumstances without lengthy adjustment to the test. The present study was geared toward the development of such a test for use at ports-of-entry for security screening of passengers.

There are several types of polygraph tests. The method most often used for screening

examinations is the relevant-irrelevant test (RIT). A typical RIT contains four relevant questions and several irrelevant questions. Relevant questions pertain to matters under investigation (e.g., “Do you intend to disrupt or interfere with a flight today?”). Irrelevant questions are interspersed among the relevant questions to provide baseline measures of autonomic reactivity when the subject is truthful to innocuous questions (e.g., “Is today Tuesday?”). The RIT predicts that a person who lies when asked a relevant question was concerned or fearful that their deception was revealed by the polygraph, and that concern was associated with a strong autonomic response. Results from laboratory and field studies confirm that prediction. The RIT also predicts that people who are truthful to all of the questions on the test will show little or no difference in the strength of responses to relevant and irrelevant questions. That assumption is problematic. Since it is obvious to anyone who takes the test that the relevant questions are more important than the irrelevant questions, the relative salience of relevant questions may cause people to react more strongly to them, whether they are deceptive or not. Consequently, when responses to relevant and irrelevant questions are compared, there is a tendency for truthful individuals to fail the test. These are known as false positive errors. Comparisons of reactions to relevant and irrelevant questions result in unacceptably high rates of false positive outcomes (Horowitz et al., 1997).

The comparison-question test (CQT) was developed to reduce the risk of false positive errors. CQTs include comparison questions that are designed to evoke relatively strong physiological responses in individuals who answer relevant questions truthfully. CQTs are commonly used in criminal investigations and to investigate indications of deception following an RIT (Krapohl, 2001). Whereas the RIT predicts that truthful individuals will show little or no difference between reactions to relevant and irrelevant questions or among reactions to various relevant questions, the CQT predicts that truthful subjects will react more strongly to comparison questions than to relevant questions. These predictions have been confirmed in

many laboratory and field studies. The accuracy of decisions from properly conducted and interpreted CQTs is approximately 90% for both truthful and deceptive individuals (Raskin & Kircher, 2014).

The current research project evaluated a new relevant-relevant comparison test format (RCT). The RCT consisted of two sets of relevant questions. One set of relevant questions addressed matters related to security, whereas the other set of relevant questions addressed matters related to the transport of illegal drugs. With the War on Terror and the War on Drugs, both issues are important to the government and both have face validity among the traveling public. Since both issues are important, we would not expect people who are truthful to respond more strongly to one set of relevant questions than to the other. In contrast, a person who answers one set of questions deceptively would be expected to respond more strongly to that set of questions than to the other set. Theoretically, the RCT could be defeated if the passenger were deceptive on both issues and reacted strongly to all of the relevant questions. However, is difficult to imagine a situation where, for example, a person would simultaneously attempt to transport drugs *and* transport a bomb. Recent research supports the hypothesis that if relevant questions address important but disparate issues, then people who are truthful to all relevant questions will not show especially strong responses to questions that address one particular issue (Brownlie et al., 2001; Honts & Amato, 2007).

Computerized polygraph test administration has delivered significantly more accurate outcomes than those administered by human polygraph examiners (Honts & Amato, 2007). Honts and Amato (2007) compared tests administered by polygraph examiners to those administered by a fully automated computer system. In both conditions, participants assigned to the guilty condition falsified information on employment questionnaires and then lied when asked if the information was accurate. For the automated condition, decisions determined by optimal cutoffs yielded 77.5% accuracy compared to 65% for human polygraph examiners.

Traditional polygraphs measure cardiovascular responses with a blood pressure cuff that is wrapped around the upper arm and inflated to about 60 mm Hg. The recording generated by the polygraph is known as a cardiograph. Research indicates that the cardiograph covaries with changes in arterial blood pressure (Podlesny & Kircher, 1999) and is diagnostic of deception in CQTs (e.g., Kircher & Raskin, 1988a; Podlesny & Kircher, 1999; Raskin et al., 1988). However, after several minutes, the cuff causes vasocongestion in the arm below the cuff, which produces discomfort.

To minimize participant discomfort, the time required to collect data, and demands on the operator, a Finometer Pro (FMS Biomedical Instrumentation, The Netherlands) was used to measure cardiovascular reactions. Without the cardiograph, the automated system was able to present a single, uninterrupted series of test questions, increase the number of times each question was asked, and reduce total test time. The Finometer Pro was used in the present study as a possible substitute for the cardiograph. The Finometer Pro is a newer version of the Finapres blood pressure monitor. Like the Finapres, the Finometer Pro monitors arterial blood pressure from a low-pressure finger cuff. In two prior experiments, the Finapres was slightly more effective than the cardiograph in discriminating between truthful and deceptive individuals (Kircher et al., 1998; Kircher et al., 2001). In another study, the Finapres was significantly more effective than the cardiograph (Podlesny & Kircher, 1999). In addition to being less invasive, prior research suggested that the Finometer Pro or similar technology would perform at least as well as the cardiograph in a screening context.

The Finometer provided continuous measures of blood pressure and beat-by-beat measures of stroke volume, cardiac output, and peripheral resistance. Since these are the physiological determinants of arterial blood pressure, one or two of the measures might be more diagnostic than blood pressure itself. The present study explored that possibility.

Recent work with CQTs indicated that increases in pupil diameter were strongly

correlated with increases in skin conductance (*r* = .57; Webb et al., 2009). That research also suggested that pupil diameter may be as useful as skin conductance for detecting deception. Skin conductance and pupil diameter correlated .62 and .61 with group membership, respectively. If similar results could be obtained with the automated RCT, skin conductance could be replaced by pupil diameter with no decrease in diagnostic accuracy. In contrast to skin conductance, pupil diameter may be measured remotely with off-the-shelf eye trackers, and measurement does not require application of surface electrodes, which would be problematic in a high volume screening environment.

Decision Algorithms

The available sample of 354 subjects was split randomly into standardization and validation samples, subject to the constraints that the validation sample was balanced with respect to group assignment (drug, bomb, or innocent) and phase. Physiological data obtained in the standardization sample were used to generate several statistical models for classification that were cross-validated in the validation sample. Bivariate correlations between each of several physiological measures and guilt status were used to assess the predictive ability of each measure. Results from discriminant analyses were compared to results from computer-intensive, machine learning methods known as bootstrap aggregating ("bagging"), and boosting.

Bootstrap aggregating ("bagging") and boosting are designed to identify predictor variables and procedures that decide if a person belongs to one group of individuals or another (e.g., bomb, drug, or innocent). Bagging and boosting techniques use a combination (or ensemble) of decision models rather than using a single model (Buhlmann, 2004). These techniques are based on the assumption that error is specific to each model, and the use of a combination of models reduces error.

Bagging was designed to reduce variance and mean squared error in a decision

algorithm (Breiman, 1996). It attempts to reduce overfitting of the data and, therefore, increase the generalizability of the classification algorithm. Bagging attempts to approximate the true population by sampling from a data set multiple times. Bagging assumes that classifiers of the data set have already been identified by their ability to correctly predict group membership and attempts to improve those classifiers. Samples of the data are drawn from the data with replacement, and these samples are used to "train" the model to produce an improved predictor of any given data point. Each sample of the data is used to produce a classification model are averaged (regression) or given majority rule (classification) to form a final classification model.

Bagging assumes that variables in the classification model are independent. In mathematical models of bagging, bagging has been found to reduce variance because the variance of the data set is divided by the number of bootstrap samples of the set (Buhlmann & Yu, 2002). Bagging also reduces the mean squared error. Generally, bagging is effective at reducing variance only when the data set is unstable, that is, when small changes in the training set results in large changes in predictions (Breiman, 1996). Essentially, bagging is effective when there is enough variance that different random samples from the population set could yield widely discrepant results. As bagging essentially operates as a "smoother" of a classification model or regression line, there has to be enough variance to smooth in order for bagging to be effective.

Boosting is another strategy to improve decision accuracy (Freund & Schapire, 1996). Boosting uses a somewhat more complicated process to develop a decision algorithm that consists of identifying a multitude of weak predictors and combining them into a single decision algorithm. Boosting starts with a data set sampled from the whole and identifies the best classifier. This classifier is often what is termed "weak" in that its accuracy is only slightly better than chance. The errors of that classifier are given greater weight in the next iteration so the

next classification model does a better job of correctly classifying those cases. This process is repeated until a classification algorithm is obtained that consists of a subset of optimally weighted predictors. Each iteration can make use of "weak learners," classification models that, by themselves, do not produce good results, but they produce a complex decision model that ultimately yields highly accurate decisions. Rather than attempting to simplify a classification model as in bagging, boosting is a method to increase the complexity of a decision model. This process tends to reduce generalizability, but it could, in theory, provide higher accuracy on cross-validation than bagging.

Research Objectives

The present study had four primary objectives:

1. To evaluate the Computerized Screening System for ports of entry
2. To evaluate a new Relevant Comparison Test (RCT) for use at ports of entry
3. To evaluate less invasive alternatives to skin conductance and the cardiograph
4. To evaluate alternative statistical methods for classification

CHAPTER 2

METHOD

Participants

Two hundred Phase 1 and 217 Phase 2 participants were recruited from the general community by newspaper and online advertisements (www.saltlakecity.craigslist.org). The advertisements offered $20 of compensation per hour with an opportunity to earn an additional $80 bonus. Of the 417 participants, 63 were eliminated from the study. Four declined to participate after being assigned to a guilty condition. Sixteen failed to follow their directions and were dismissed after being compensated for their time. Another 43 participants produced unusable data due to equipment failure. The data for the remaining 354 participants were evaluated.

A majority (58.8%) of participants were male. The mean age of the participants was 31.0 (SD = 12.5). Years of education ranged from 10 (10^{th} grade of high school) to 21 years (M = 14.7; SD = 3.3). Most participants were Caucasian (78.5%), followed by Hispanic (9.0%), Pacific Islander (2%), African-American (2%), and Native American (2%). The remaining 6.1% identified as 'Other.' Participants estimated the number of domestic flights and the number of international flights they took in the past year. The number of domestic flights ranged from 0 to 200 (M = 5.26, SD = 13.7); international flights ranged from 0 to 30 (M = 0.81, SD = 2.3).

Procedures

Prospective participants called a research assistant who screened callers for eligibility and briefly described the experiment and pay policy. Callers were invited to participate if they met the following criteria: (1) they were between 18 and 65, (2) they were not taking prescription medication, (3) they had never had a polygraph test, (4) they were fluent in English, and (5) they had no major medical problems.

Callers who agreed to participate were given an appointment to report to a room in a building on the campus of the University of Utah. Participants received a reminder phone call the day before their appointment. When the participant arrived, an envelope addressed to the participant was taped to the door. Instructions within the envelope directed the participant to enter the room, close the door, read and sign an informed consent form, complete a brief questionnaire, and then play a cassette recorder that presented their instructions over headphones.

Participants were randomly assigned to bomb, drug, or innocent conditions according to a predefined running order. One-hundred eleven participants were assigned to the bomb condition, 119 were assigned to the drug condition, and the remaining 124 were innocent.

Guilty participants received tape-recorded instructions to transport a yellow gym bag containing a jacket and a cell phone to a room two floors below. The participants were instructed to use the set of keys provided to them to open the door to the room and open a locked cabinet within. They were then expected to look through several boxes in the cabinet to find a 'bomb' or what appeared to be illegal drugs, depending on the condition to which they were assigned. For the bomb condition, participants were expected to connect two wires to illuminate a small flashing bulb on the apparatus. For the drug condition, participants were instructed to place several small plastic bags containing a white powder into a larger bag. For both conditions, participants were instructed to conceal either the bomb or the drugs in a

hidden compartment in a yellow gym bag, lock the cabinet and room, and then present themselves to the laboratory. Innocent participants listened to a general description of the crime, left the area for 15 min, and then went to the laboratory where they met the research assistant. The research assistant searched the yellow bag and asked participants if they had transported anything illegal. Participants were dismissed if they admitted to transporting something illegal when asked by the research assistant, since they had been instructed to deny taking and transporting anything illegal in the bag.

The participant then was instructed to go to the restroom and wash their hands with soap and warm water. The test was administered when the participant returned from the restroom. The session was videotaped and audiotaped.

Participants were told that physiological measures would be recorded as they were asked questions about a crime. They were told that the computer program would determine if participants had been truthful or deceptive on the test, and that if the participants could convince the computer program of their innocence, they would receive $20 in pay plus a bonus payment of $80. They also were told that if they appeared deceptive on the test, they would receive only $20 per hour for their time.

The research assistant obtained biographical information from the participants and asked some questions about their health (Appendix A). Participants who reported less than 6 hr of sleep, were experiencing pain, or indicated that they had recently taken stimulant or depressant drugs were not tested; they were paid for their time and released. For the remaining participants, the sensors were attached. The research assistant explained that they would hear some instructions and then be asked a series of test questions over a loudspeaker in the testing room. They were told to answer each question Yes or No and avoid any unnecessary movements once the test began. The research assistant asked if they had any questions, then left the room and left the door ajar.

The Finometer was then calibrated according to manufacturer's instructions. This was followed by verbal presentation by the computer of recorded instructions and test questions over a loudspeaker which was placed on a small table next to the participant.

In Phase 1, a computer-generated voice presented the instructions and test questions (AT&T Voices, Kate16). In addition, the computer evaluated the respiration and finger photoplethysmograph data as they were being collected. If a large breath or high frequency, high amplitude artifact was detected in the finger plethysmograph, the computer automatically issued a cautionary statement ("Abnormal breathing will lengthen the test. Please breathe normally;" or "Please remain still. Movements will lengthen the test"). In Phase 2, a pre-recorded female human voice presented the instructions and test questions, and no cautionary statement was provided if a deep breath or movement was detected. Computer administered instructions are included in Appendix B. It required 3.4 min to present the instructions and 15.5 min to present the test questions. All procedures were approved by the University of Utah Internal Review Board (IRB) and the U.S. Army IRB (USAMRMC) prior to data collection.

Relevant Comparison Test (RCT)

The RCT consisted of the following 16 test questions, though not in this order:

	Question type	Question
1	Relevant	Drugs: Did you take illegal drugs from a locked cabinet?
2	Relevant	Drugs: Did you put illegal drugs in a flight bag?
3	Relevant	Drugs: Did you bring illegal drugs into this room?
4	Relevant	Drugs: Are there illegal drugs hidden in the flight bag?
5	Relevant	Bomb: Did you take a bomb from a locked cabinet?
6	Relevant	Bomb: Did you put a bomb in a flight bag?
7	Relevant	Bomb: Did you bring a bomb into this room?
8	Relevant	Bomb: Is there a bomb hidden in the flight bag?
9	Neutral	Are you sitting down?
10	Neutral	Is this the year 2009?
11	Neutral	Is this the year 1996?
12	Neutral	Are you in Seattle, Washington?
13	Neutral	Have you ever been to Salt Lake City, Utah?
14	Neutral	Is today Sunday?

15	Neutral	Are you older than 16 years of age?
16	Neutral	Are the lights on in this room?

In Phase 1, each relevant question that addressed drugs began with the word “Drugs,” and each question that addressed the explosive device began with the word “Bomb.” The words “Drugs” and “Bomb” were prepended to the relevant questions to distinguish the relevant issue at the outset of the question and reduce the similarity in the wording of the relevant questions that addressed the two crimes. The pretest instructions prepared Phase 1 participants for this format. In Phase 2, the words “Drugs” and “Bomb” were not stated at the beginning of each relevant question.

The series began with two neutral questions. Physiological reactions to the first neutral question were not evaluated due to orienting reactions. Relevant and neutral questions were then alternated in the question sequence. A neutral question was presented after each relevant question to give a large reaction to a relevant question an opportunity to recover before another relevant question was presented. To minimize the duration of the test, physiological data were collected for 16 s following the onset of each neutral question and 22 s following the onset of each relevant question. Within relevant and neutral categories, the order of presentation was randomized. The set of 16 test items was repeated three times in different orders. Altogether, 49 test questions were presented (16 questions x 3 repetitions + 1 initial neutral question), reactions to the last 48 of which were evaluated.

Apparatus

In Phase 1, the CPS-II (Stoelting, Wood Dale, IL) recorded thoracic respiration, skin conductance, and finger pulse amplitude at 60 Hz. In Phase 2, the CPS-Pro (Stoelting, Wood Dale, IL) recorded respiration, skin conductance, and finger pulse amplitude data at 60 Hz. Respiration was recorded from a Pneumotrace transducer (UFI, Morro Bay, CA) secured with Velcro around the upper chest. Skin conductance (SC) was obtained by applying a constant

voltage of .5V to two disposable Ag-AgCl snap electrodes attached to the distal phalanx of the ring and last fingers of the left hand. Finger pulse amplitude (FPA) was obtained from a UFI photoplethysmograph attached to the first finger of the left hand with a Velcro strap.

Four calibrated analog outputs from the Finometer Pro (FMS, Finapres Medical Systems, BV, Amsterdam, The Netherlands) were digitized at 60 Hz by a Measurement Computing 16-bit model 1608 USB analog-to-digital converter (Norton, MA). The Finometer arm cuff was secured around the left upper arm and the finger cuff was placed on the middle phalanx of the left middle finger. The continuous analog outputs provided by the Finometer included brachial arterial pressure (BP), total peripheral resistance (TPR), left ventricular ejection time (LVET), and stroke volume (SV).

In Phase 1, an AcuNetx Physiological Detection of Deception System (AcuNetx, Torrance, CA) was used to record changes in right pupil size at 60 Hz. A camera and infrared light source was mounted on an open face mask that rested on the participant's face during the test. The face mask was mounted on an articulated arm attached to the chair. The chair was reclined by approximately 15% to allow the mask to rest comfortably against the face during the test.

In Phase 2, an Arrington EyeFrame system was used to record changes in right pupil size at 60 Hz. A camera and infrared light was mounted on a pair of plastic lensless plastic goggles worn by the participant during the test. The participant adjusted a strap attached to the temples of the goggles around the back of the head to ensure that they remained snugly in place and the distance between the camera and the eye remained constant during the test. Some participants in Phase 1 appeared to get drowsy during the test. In an effort to counter sleepiness, the reclining chair used in Phase 1 was replaced by a standard, lightly cushioned office chair that rested on four legs, did not recline or rotate, and required the participant to sit upright with both feet on the floor during the test. The chair was refitted with oversized custom wooden arm rests.

Summary of Differences between Phase 1 and Phase 2

Table 1 summarizes the procedural differences between Phase 1 and Phase 2.

Computer Measurements of Physiological Waveforms

CPSLAB (Scientific Assessment Technologies, Homer, AK) was used to extract waveforms and features from the physiological signals.

Respiration

The 60 Hz respiration signal for each participant was transformed to a set of z scores to adjust for arbitrary changes in gain and offset. Respiration line length was a sum of absolute deviations between successive 60 Hz respiration samples. The 60 absolute difference scores for each second was summed to obtain a second-by-second respiration waveform that began at question onset and ended 16 s later.

Skin Conductance (SC)

The 60 Hz skin conductance signal for each participant was transformed to a set of z scores to adjust for arbitrary changes in gain and offset. Starting at the onset of each test

Table 1. *Procedural differences between Phase 1 and Phase 2.*

	Phase 1	**Phase 2**
Voice	Computer-generated female	Digitally recorded human female
Relevant questions	Preceded by the word “Bomb” or “Drugs”	Not preceded by “Bomb” or “Drugs”
Automated feedback	The computer provided feedback if an artifact was detected. Relevant questions were repeated if corrupted by artifact.	No feedback was provided. Relevant questions were not repeated if corrupted by artifact.
Polygraph chair	Recliner	Wood frame, nonreclining
Physiological recorder	CPS-II (Stoelting Company, Wood Dale, IL)	CPS Pro (Stoelting Company, Wood Dale, IL)
Eye tracker	HawkEye (AcuNetx, Torrence, CA)	Arrington EyeFrame (Arrington Research, Scottsdale AZ)

question, the mean of each successive set of 30 samples was computed to reduce the 60 Hz samples of SC to 2 Hz for 16 poststimulus seconds. A .5-second-by-.5-second SC waveform was defined by subtracting the value for the first second from each poststimulus second.

Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP)

CPSLAB identified the times and levels of systolic and diastolic points in the 60 Hz blood pressure and vasomotor signals that began at question onset and ended 16 s later. It calculated a weighted average of the systolic points that occurred during each poststimulus second starting at question onset (Kircher & Raskin, 2001). The resulting series of 16 poststimulus systolic averages defined a systolic waveform (SBP). The same procedure was used to create a second-by-second diastolic waveform (DBP).

Left Ventricular Ejection Time (LVET)

CPSLAB sampled the LVET signal generated by the Finometer at 60 Hz. Starting at the onset of each test question, the mean of each successive set of 60 samples was computed to reduce the 60 Hz samples of LVET to 1 Hz for 16 s following question onset. A second-by-second LVET waveform was defined by subtracting the value for the first second from each second following question onset.

Stroke Volume (SV)

CPSLAB processed the SV signal generated by the Finometer in the manner described above for LVET.

Total Peripheral Resistance (TPR)

CPSLAB processed the TPR signal generated by the Finometer in the manner described above for LVET.

Finger Pulse Amplitude (FPA)

The 60 Hz vasomotor signal for each participant was transformed to a set of z scores to adjust for arbitrary changes in gain and offset. Separate second-by-second reaction waveforms were obtained for systolic and diastolic points in the manner described above for blood pressure. Second-by-second changes in FPA were obtained by subtracting the diastolic level for each second from the corresponding systolic level. The 16 poststimulus pulse amplitudes were expressed as proportions of initial value by dividing each of the 16 poststimulus amplitudes by the first value following question onset.

Heart Rate (HR)

Heart rate was obtained from the times between systolic points in the blood pressure signal measured to the nearest 17 ms. Second-by-second interbeat interval (IBI) waveforms were obtained by computing a weighted average of the times that occurred during each post-stimulus second. The IBIs in milliseconds was divided into 60,000 to obtain HR in beats per minute.

Pupil Diameter (PD)

Recordings of pupil size by eye trackers are characterized by relatively slow baseline changes in pupil diameter that are interrupted by occasional eye blinks. Because the eye is closed momentarily during the blink, the eye tracker loses the pupil and has to reacquire it when the eye opens. The recorded signal shows high frequency, high amplitude changes that would contaminate the reaction waveform if they were not removed. A modified version of an algorithm recommended by Marchak and Tanner (personal communication, 2009) was used to detect eye blinks or other losses in the pupil signal. First, the pupil signal was converted to a set of z scores. Any z score that exceeds +/-4 was considered an outlier and was replaced with a missing value code. Second, the pupil data again was transformed to z scores except that any

previously identified outlier was omitted from the standardization process. The new set of z scores again was tested for outliers; in effect, z scores that exceed +/-4 were replaced with missing value codes. This procedure was repeated a total of five times. After the fifth iteration, the computer interpolated across the regions of the signal marked as missing.

To obtain a PD reaction waveform, the edited 60 Hz pupil diameter signal for each participant was transformed to z scores and reduced to 1 Hz by computing the mean for each successive group of 60 samples for 16 or 22 s. A second-by-second PD waveform was defined by subtracting the value for the first second from each poststimulus second.

Feature Extraction

Respiration Reactions

Respiration line length (RLL) was extracted for 10 s starting at question onset. This scoring window for RLL was based on prior research with comparison question test formats (CQT; Kircher & Raskin, 2001).

SC, Cardiovascular, and Pupil Reactions

Peak amplitude and area under the curve were extracted from SC, cardiovascular, and pupil waveforms. Since a reaction in the FPA channel was indicated by a reduction in amplitude (vasoconstriction), each value in the FPA waveform was reflected (multiplied by -1) prior to feature extraction. For FPA, area under the curve and duration of the reaction was extracted from the reflected second-by-second FPA waveform.

Amplitude

Low points in the waveform were identified as changes from negative or zero slope to positive slope, and high points in the waveform were identified as changes from positive slope to zero or negative slope. The difference was measured between each low point and every

subsequent high point. Amplitude was defined as the greatest observed difference.

Area Under the Curve (AUC)

The level of the low point at which peak amplitude was measured was subtracted from each subsequent value in the waveform until the level dropped to the initial baseline value or to the end of the 16-s scoring window, whichever came first. AUC was the sum of the differences.

Duration of the FPA Reaction

The time from the low point at which peak amplitude was measured was subtracted from the time at which the reaction returned to the initial pulse amplitude or 16 s, whichever occurred first.

HR Reactions

HR reactions in deception detection studies typically are characterized by an initial acceleration followed by a larger deceleration and a gradual return to prestimulus levels (Raskin & Hare, 1978). The following features were extracted from the HR waveform.

Amplitude of the Initial Increase in HR

The algorithm described above for SC amplitude was used to measure the greatest increase in HR in the first 8 s following question onset.

Amplitude of the Decrease in HR

To measure a decrease in HR that follows an initial increase, the second-by-second HR waveform was reflected such that a deceleration was indicated by a rise in the waveform. The amplitude algorithm described above was used to measure the amplitude of the HR deceleration waveform in the 16 s following question onset.

Within-subject Standardization of Physiological Features

Each participant answered 12 questions concerning drugs, 12 questions concerning the bomb, and 24 questions concerning neutral (irrelevant) topics. The 48 measurements of a particular physiological feature (e.g., SC amplitude) were converted to z scores. The mean of the measurements for the 12 drug questions were computed. A mean also was obtained for the 12 bomb questions, and another mean was obtained for the 24 neutral questions.

Indices of Differential Reactivity

The RCT contained relevant questions that addressed two relevant issues (drugs and bomb). The RCT predicted that a person who was deceptive about transporting drugs would react more strongly to questions about the drugs, whereas a person who was deceptive about transporting a bomb would react more strongly to questions about the bomb. An innocent person who answered the relevant questions truthfully should react similarly to the two sets of relevant questions.

Two indices of differential reactivity were computed for each of the physiological features described above. The indices of differential reactivity were used to create statistical classifiers that sorted participants into innocent, drug, and bomb groups based on their reactions to the three types of test questions (neutral, drug, and bomb). One index of differential reactivity was the difference between the means for drug (R1) and bomb (R2) relevant questions. I expected the (R1-R2) difference to be positive if the participant was deceptive to the drug questions, negative if the participant was deceptive to the bomb questions, and near zero if the participant was truthful to drug and bomb questions.

To obtain the second index of differential reactivity, the mean for neutral questions (N) was subtracted from the mean for all drug (R1) and bomb (R2) relevant questions combined; that is, ((R1 + R2)/2) – N. This difference was not expected to be as diagnostic of group

membership as the difference between reactions to drug and bomb questions. However, this difference was orthogonal to the (R1-R2) comparison and provided an independent source of information to distinguish guilty from innocent participants.

For all variables except respiration, a large measured reaction was indicative of a strong reaction. For RLL, suppressed respiratory activity was indicative of a strong reaction. Participants guilty of transporting drugs are expected to show relatively small measures of respiratory activity (suppression) to drug questions, whereas participants guilty of transporting the bomb was expected to show relatively small measures of respiratory activity (suppression) to bomb questions. To achieve a common direction for predicted effects, the sign of the index of differential reactivity for respiration was reversed.

Classification Techniques

Discriminant functions were developed from the data in the standardization sample and were used to classify cases in the validation sample. This approach to selecting and weighing variables for classification problems is well established and commonplace in the literature (Kircher & Raskin (1988b; 2001). Newer, computer intensive methods are now available for classification problems that often out-perform traditional approaches such as discriminant analysis and logistic regression analysis (Breiman, 1996; Freund & Schapire, 1996). The current study compared outcomes from discriminant analysis to two computer intensive procedures: bagging and boosting.

In the current study, the predictor variables were the various physiological measures of arousal and the outcome was categorical membership in one of three treatment conditions: bomb, drug, or innocent. Bagging and boosting functions for a continuous outcome variable are available in statistical analysis packages. For the present multinomial classification problem, the statistical package AdaBag (R Core Team, 2015) was used.

CHAPTER 3

RESULTS

Preliminary Analysis: Effects of Procedural Changes between Phase 1 and Phase 2

A preliminary analysis was conducted to determine if procedural changes made between Phase 1 and Phase 2 affected the diagnostic value of any of the physiological features. If not, the data for Phase 1 and Phase 2 could be pooled for many of the analyses to avoid redundancy. On the other hand, if there were large differences in the diagnostic utility of the physiological measures, it would be necessary to report the results for Phase 1 and Phase 2 separately.

For each physiological feature, repeated measures analyses of variance (RMANOVA) were conducted with three factors. Condition was a between-group factor with three levels (drug, bomb, and innocent). Question Type was a within-subject factor with three levels (neutral, drug, and bomb questions). Phase was a between-group factor with two levels (Phase 1 and Phase 2). Huynd-Feldt corrected degrees of freedom were used to reduce the numbers of degrees of freedom for tests involving more than two levels of a repeated factor (i.e., Question Type). In RMANOVA, the diagnostic usefulness of each physiological feature was indicated by the Condition X Question Type interaction. To determine if the diagnostic utility of the feature varied between phases, we tested for the presence of a Condition X Question Type X Phase interaction. The results from the RMANOVAs are presented in Table 2.

Table 2. *Effects of changes from Phase 1 to Phase 2 on physiological features: Condition X Question Type X Phase interaction effects*

Signal	Feature	Df	F	P	Partial η^2
Thoracic respiration	Line length 0-10s	(4, 696)	2.74	.028	.016
	Line Length 6-12s	(4, 696)	1.89	-	-
Skin conductance	Peak amplitude	(4, 696)	1.13	-	-
	Area under the curve	(4, 696)	.44	-	-
Finger pulse amplitude	Area under the curve	(4, 696)	.79	-	-
	Duration	(4, 692.4)	1.98	-	-
Systolic blood pressure	Peak amplitude	(4, 696)	.86	-	-
	Area under the curve	(4, 696)	1.11	-	-
Diastolic blood pressure	Peak amplitude	(4, 696)	1.26	-	-
	Area under the curve	(4, 696)	1.33	-	-
Left ventricular ejection time	Peak amplitude	(4, 696)	1.73	-	-
	Area under the curve	(4, 696)	.22	-	-
Stroke volume	Peak amplitude	(4, 696)	2.99	.02	.018
	Area under the curve	(4, 696)	2.11	-	-
Total peripheral resistance	Peak amplitude	(4, 696)	2.89	.02	.018
	Area under the curve	(4, 696)	2.67	.03	.016
Heart rate	Maximum increase 0-8s	(3.9, 683.7)	.69	-	-
	Maximum decrease 0-16s	(4, 696)	2.16	-	-
Pupil diameter	Peak amplitude	(4, 681.4)	2.26	-	-
	Area under the curve	(4, 696)	1.01	-	-

RMANOVA revealed a small effect of Phase on the diagnostic usefulness of thoracic respiration line length (RLL) for the 0-10s interval (partial η^2 = .016) and three of 14 cardiovascular measures (SV peak amplitude, TPR peak amplitude, and TPR area under the curve). To determine which of the two phases provided better data, the Condition X Question Type interaction was computed separately for Phase 1 and Phase 2 participants. This analysis indicated that the effect size for RLL in Phase 2 (Partial η^2 = .096) was twice as large as in Phase 1 (Partial η^2 = .042). The diagnostic effects for SV peak amplitude and TPR area under the curve were also greater in Phase 2 than in Phase 1. For TPR area under the curve, the Condition X Question Type effects were stronger in Phase 1 than in Phase 2.

The results for TPR area under the curve were better in Phase 1 than in Phase 2, whereas the results for TPR amplitude were better in Phase 2. Except for TPR area under the

curve, Phase 2 provided more diagnostic respiration and cardiovascular measures than did Phase 1. These results probably were a consequence of postural changes associated with different seating arrangements. Phase 1 participants reclined slightly in a cushioned recliner, whereas Phase 2 participants sat upright in a rigid wooden chair with arm rests. The latter arrangement appeared to improve the usefulness of the respiration, SV, and one of the two TPR features. However, all of the observed effects were small. None accounted for even 2% of the variance in the physiological measure.

Because the effects of Phase were negligible, Phase 1 and Phase 2 participants were analyzed together until variables were selected, weighed, and combined for a decision model. At that point, only participants in the standardization sample were used to develop the statistical decision models for classifying cases in the validation sample.

Analysis of Physiological Activity

Second-by-second respiration, skin conductance, cardiovascular activity, and pupil size were analyzed separately with RMANOVA. Each RMANOVA had two between-group factors. One between-group factor was condition with three levels (guilty drug, guilty bomb, and innocent). The other between-group factor was sex with two levels (male and female). The RMANOVA also had three within-subject factors: question type with three levels (drug, bomb, and neutral); block with three levels; and time with 16 levels (seconds following question onset). Each block of questions contained a complete repetition of the 16 test questions (4 drug questions, 4 bomb questions, and 8 neutrals questions). Question order varied over the three repetitions, and a neutral question always followed a relevant question. The design of the present study allowed for statistical tests for effects of repeated exposure to test questions, which could have beneficial effects (sensitization or familiarity) or adverse effects on the diagnostic validity of physiological measures (habituation or fatigue).

The RMANOVA provided statistical tests of many sources of variance, most of which were not of interest in the present study. Of interest were main effects of Condition and interactions that included Condition as a factor. We were most interested in the Condition X Question Type and the Condition X Question Type X Time interactions.

Table 3 reports the relevant results of statistical analysis of the physiological reactions. The proportion of variance in the physiological reaction explained by condition or its interaction with question type and time (partial η^2) is also reported for significant effects ($p < .05$). Figure 1 through Figure 8 present the corresponding mean second-by-second respiration, SC, cardiovascular activity (FPA, SBP, DBP, TPR, HR) and pupil size waveforms for each group and question type.

There were no significant Condition X Question Type X Sex interactions; that is, the ability of the various physiological measures to discriminate among the three groups did not differ significantly for male and female participants.

Second-by-second plots were created for physiological reaction waveforms when the Condition X Question Type interaction was significant. The figure number for signals with significant interactions is listed in the last column of Table 2. Although the Condition X Question Type interaction was not significant ($p < .08$) for TPR, it also was plotted because there was a portion of the waveform where it appeared that the three groups differed in their reactions to the three types of questions.

Examination of the second-by-second RLL data in Figure 1 reveals that the drug group showed a reduction in RLL when they lied in reaction to questions about drugs. The reaction began 6 s after question onset and lasted 6 s. The bomb group showed suppression of respiration to questions about both the bomb and the drugs. We did not predict respiration suppression to questions about drugs by participants who transported the bomb.

Figure 2 shows the change in SC for 16 s following question onset. SC reactions were

Table 3. *Repeated measures analysis of variance results for Phase 1 and Phase 2 combined*

Signal	Effect	Df	F	p	Partial h^2	Figure
Thoracic respiration	Condition	(2, 348)	1.19	-	-	1
	Condition X Question Type	(4, 696)	5.68	.001	.032	
	Condition X Question Type X Time	(37.43, 6475.96)	1.50	.025	.009	
	Condition X Question Type X Phase	(4, 696)	2.74	.028	.016	
Skin conductance[a]	Condition	(2, 348)	9.36	.001	.051	2
	Condition X Question Type	(4, 696)	54.51	.001	.239	
	Condition X Question Type X Time	(12.51,2176.6)	28.11	.001	.139	
Finger pulse amplitude	Condition	(2,348)	1.71	-	-	3
	Condition X Question Type	(3.11, 541.87)	5.46	.001	.030	
	Condition X Question Type X Time	(5.61, 976.83)	4.39	.001	.024	
Systolic blood pressure	Condition	(2, 348)	1.31	-	-	4
	Condition X Question Type	(4, 696)	41.79	.001	.194	
	Condition X Question Type X Time	(16.14, 2807.51)	18.49	.001	.096	
Diastolic blood pressure	Condition	(2, 348)	4.71	.010	.026	5
	Condition X Question Type	(4, 696)	36.55	.001	.174	
	Condition X Question Type X Time	(16.45, 2862.31)	7.33	.001	.040	
Left ventricular ejection time	Condition	(2, 348)	<1	-	-	
	Condition X Question Type	(3.89, 676.87)	1.20	-	-	
	Condition X Question Type X Time	(27.53, 4790.59)	1.28	-	-	
Stroke volume	Condition	(2, 348)	2.08	-	-	
	Condition X Question Type	(4, 696)	1.73	-	-	
	Condition X Question Type X Time	(32.56, 5666.27)	1.27	-	-	
Total peripheral resistance	Condition	(2, 348)	2.39	-	-	6
	Condition X Question Type	(4, 696)	2.14	-	-	
	Condition X Question Type X Time	(26.54, 4618.07)	1.32	-	-	
Heart rate	Condition	(2, 348)	15.75	.001	.083	7
	Condition X Question Type	(3.93, 683.3)	9.63	.001	.052	
	Condition X Question Type X Time	(26.54, 4618.07)	4.18	.001	.023	
Pupil diameter	Condition	(2, 347)	>1	-	-	8
	Condition X Question Type	(4, 696)	6.04	.001	.034	
	Condition X Question Type X Time	(27.05, 4692.42)	2.83	.001	.016	

[a]The Condition X Question Type X Repetition interaction was significant, $F(8, 1392)=4.46$, $p<.001$, Partial $\eta^2= .025$. The Condition X Question Type effect size was larger for the first repetition (Partial $\eta^2=.237$) than for the second (Partial $\eta^2=.089$) or third repetition (Partial $\eta^2=.101$).

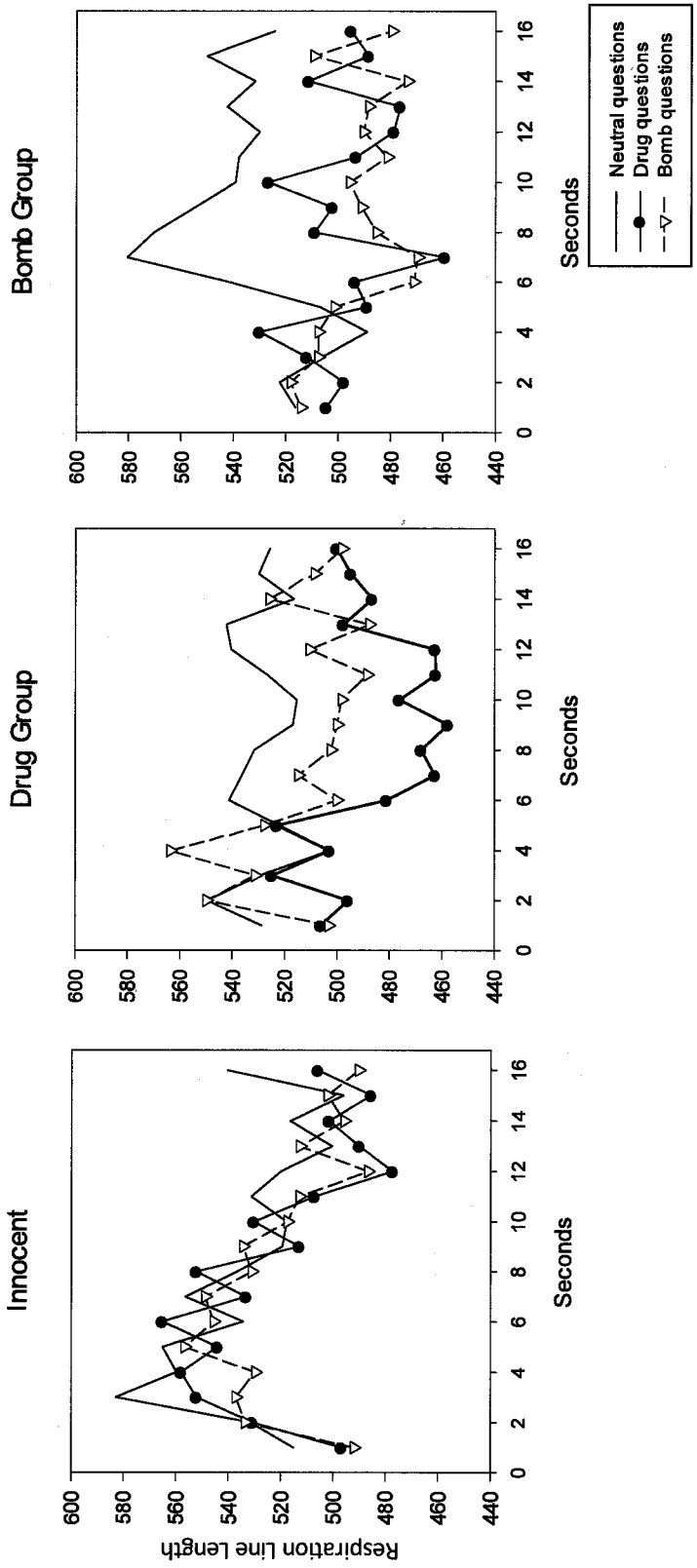

Figure 1. Mean second-by-second respiration line length for innocent, drug, and bomb groups

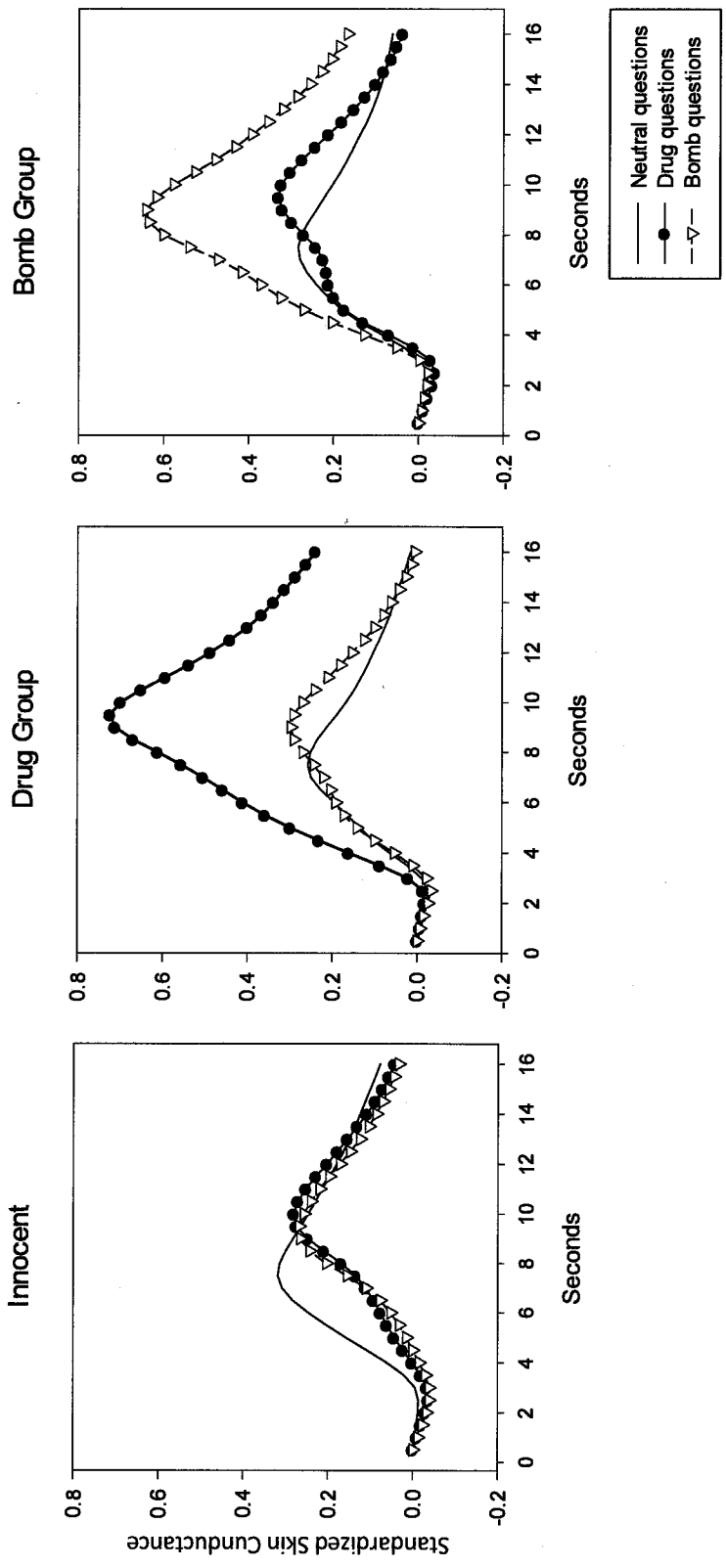

Figure 2. Mean second-by-second change in standardized skin conductance for innocent, drug, and bomb groups

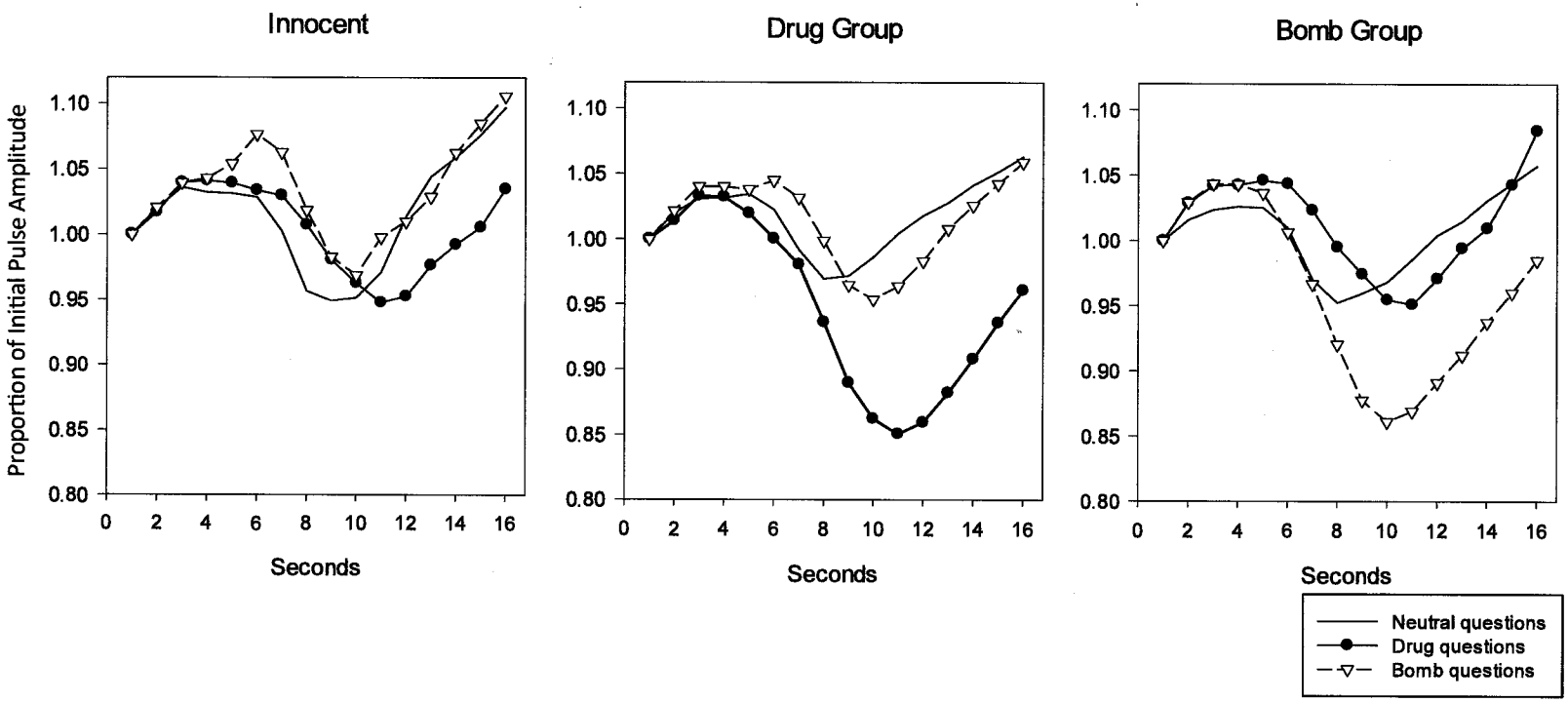

Figure 3. Mean second-by-second change in finger pulse amplitude for innocent, drug, and bomb groups

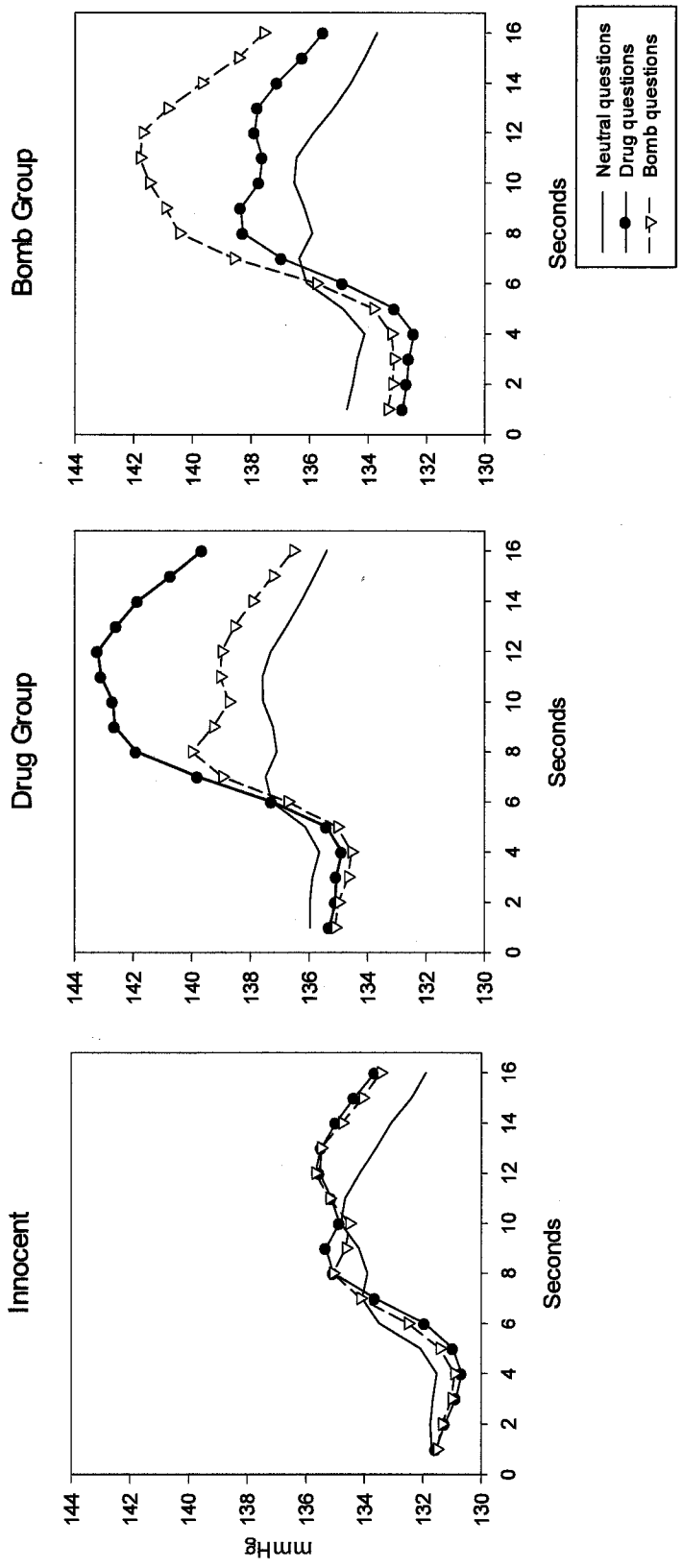

Figure 4. Mean second-by-second systolic blood pressure for innocent, drug, and bomb groups

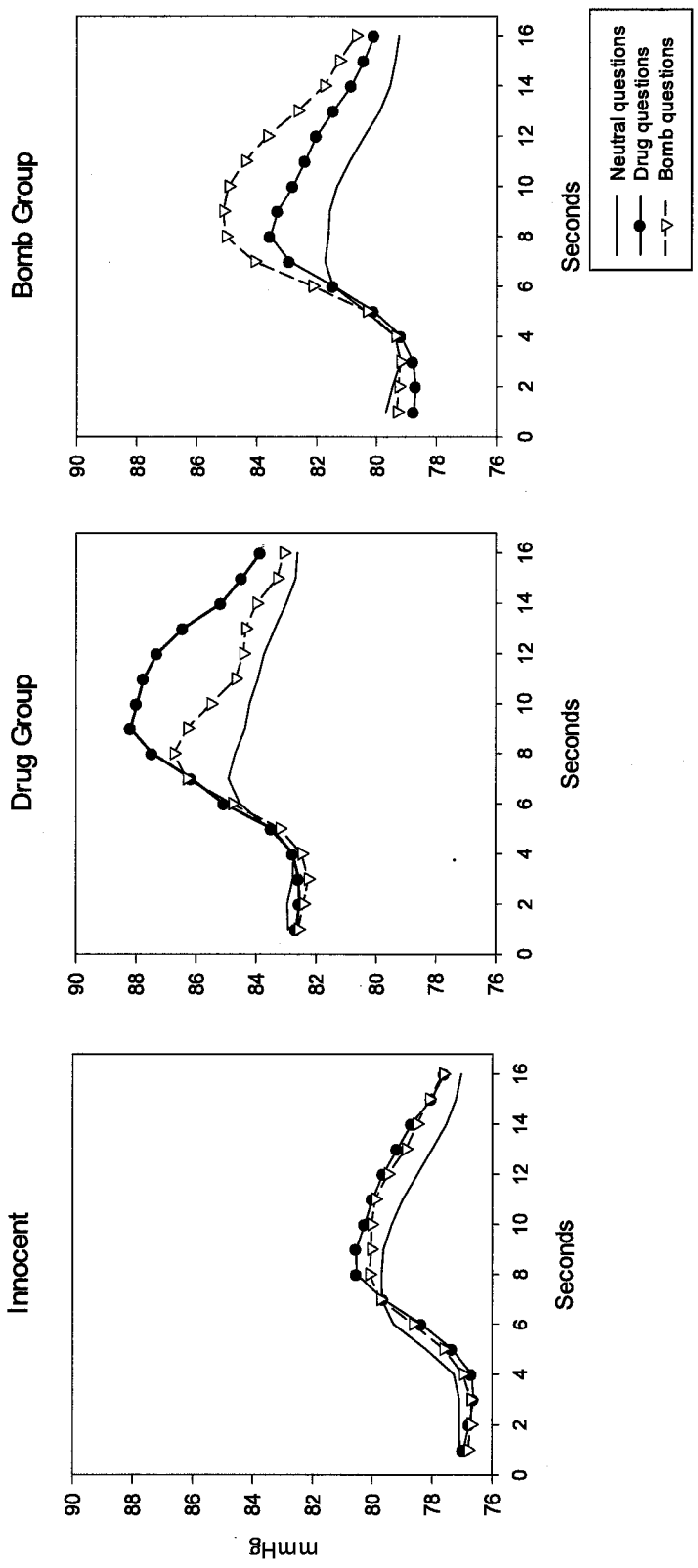

Figure 5. Mean second-by-second diastolic blood pressure for innocent, drug, and bomb groups

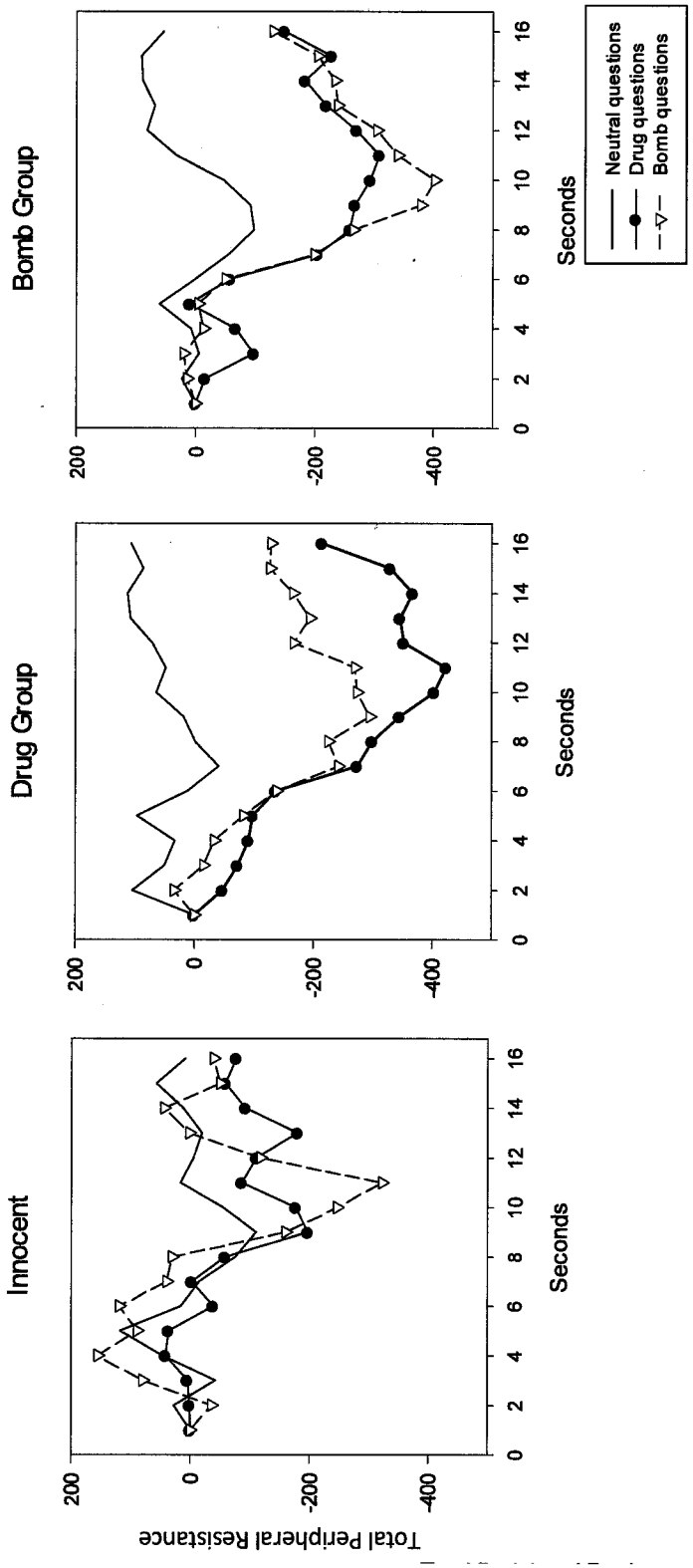

Figure 6. Mean second-by-second change in total peripheral resistance (TPR) for innocent, drug, and bomb groups

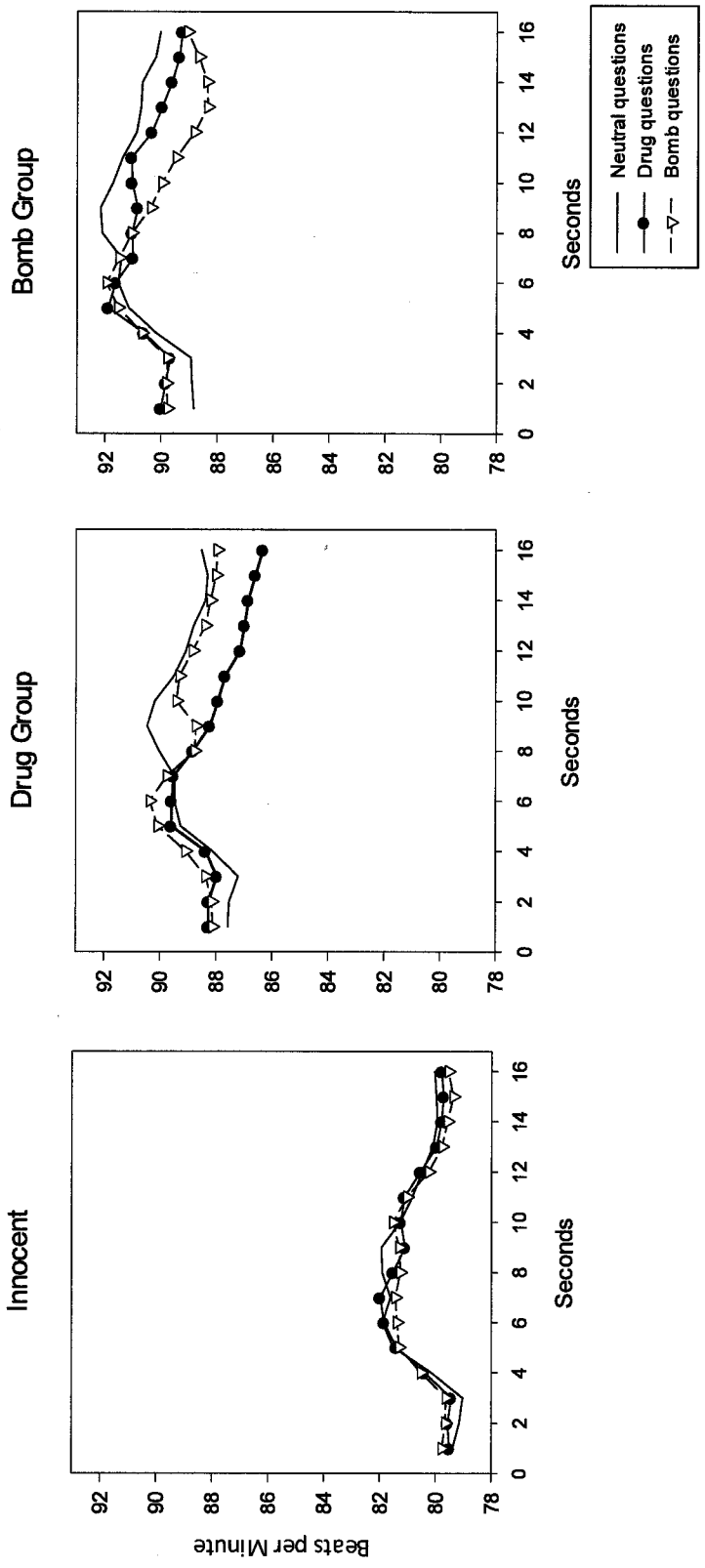

Figure 7. Mean second-by-second heart rate for innocent, drug, and bomb groups

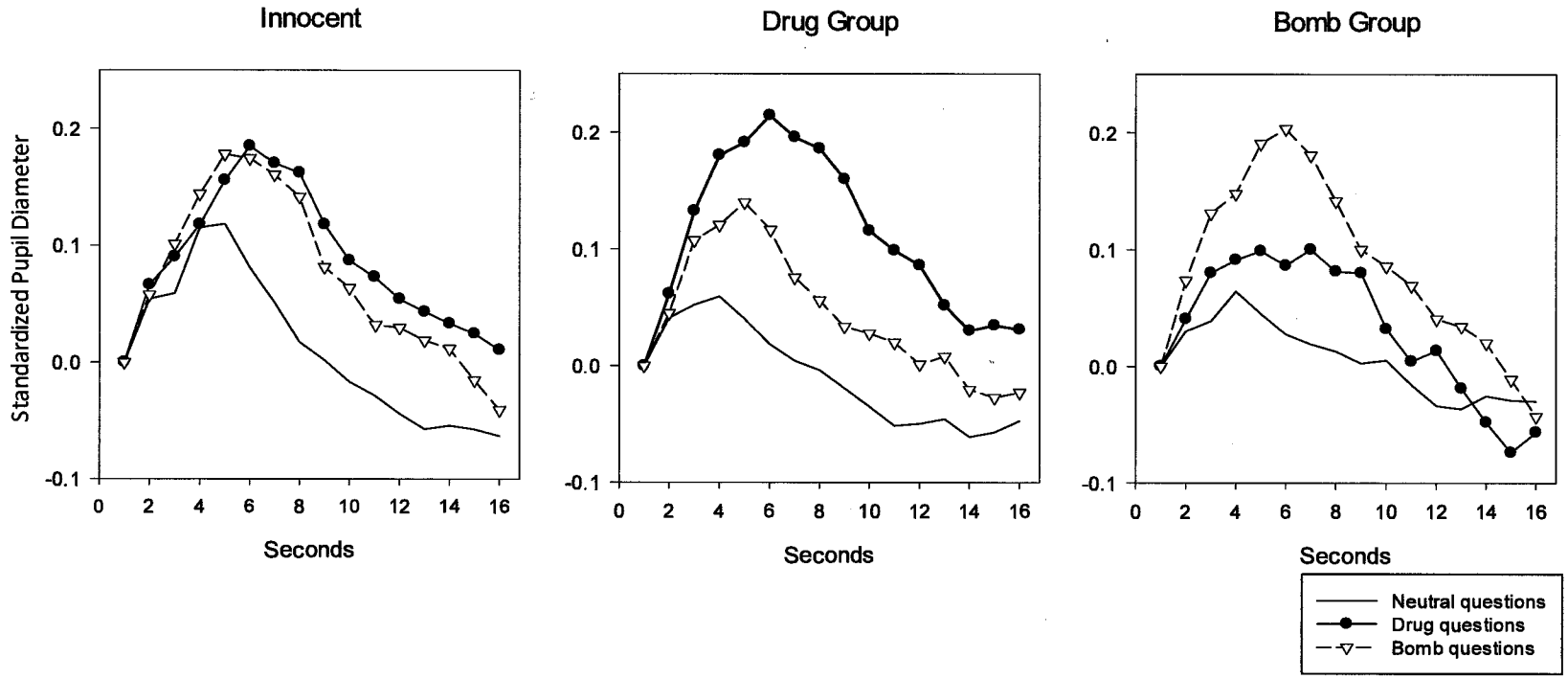

Figure 8. Mean second-by-second change in pupil diameter for innocent, drug, and bomb groups

consistent with predictions. Innocent participants showed little difference in their reactions to the three types of questions, whereas drug and bomb participants reacted more strongly to questions about the crime they had committed.

Second-by-second FPA is presented in Figure 3. The vasomotor changes were consistent with predictions. Whereas innocent participants did not react strongly or differentially to the three question types, drug participants showed more peripheral vasoconstriction to drug questions, and bomb participants showed more vasoconstriction to bomb questions.

Changes in systolic and diastolic blood pressure are shown in Figure 4 and Figure 5. Again, the blood pressure results were consistent with predictions. Within subjects, deceptive answers to relevant questions were associated with greater increases in systolic and diastolic blood pressure than were truthful answers to relevant questions. Effects of deception were greater for SBP (partial η^2 = .13) than for DBP (partial η^2 = .06).

Second-by-second TPR is plotted in Figure 6. The Condition X Question Type interaction was not significant for TPR. However, for seconds 9 through 12, TPR appeared to be lower when participants were deceptive than when they were truthful. A post hoc test of the 10-s interval between poststimulus seconds 7 and 16 revealed a significant Question Type X Condition interaction, $F(4, 696) = 2.48$, $p < .03$.

Heart rates for the three treatment conditions are presented in Figure 7. The main effect of condition was significant; heart rates were higher for the drug (M = 88.6 BPM) and bomb groups (M = 90.4) than for the innocent group (M = 80.5). There were also significant Condition X Question Type and Condition X Question Type X Time interactions, although the main effect makes it difficult to see the interactions. Careful examination of the plots for the guilty groups indicates that the recovery following an initial increase in HR was more precipitous when guilty participants were deceptive than when they were truthful.

Figure 8 displays the second-by-second pupil data for 16 s following question onset. The

effects of deception were evident in the PD waveform. As expected, the drug group reacted more strongly to questions about the drugs, and the bomb group reacted more strongly to questions about the bomb. The magnitude of the reaction associated with deception was approximately four times greater for SC (.8 SD units) than for PD (.2 SD), and the effect size as measured by partial η^2 was approximately seven times greater for SC (.239) than for PD (.034).

Bivariate Analyses

Bivariate analyses were conducted to assess the degree to which indices of differential reactivity could differentiate among the three treatment conditions. To reduce the number of variables that might be included in the statistical classifier, features were extracted from only those measures of physiological activity that significantly differentiated among the three treatment conditions. Indices of differential reactivity were obtained for respiration, SC, FPA, SBP, DBP, HR, and PD signals. Indices of differential reactivity were not obtained for LVET, SV, or TPR.

Dependent Variables

For each outcome measure, reactions to neutral, drug, and bomb questions were used to derive two indices of differential reactivity. One index of differential reactivity was the difference between reactions to drug (R1) and bomb (R2) statements (i.e., R1 – R2). We expected this difference to differentiate between the two guilty groups. Drug participants were expected to react more strongly to drug (R1) statements than bomb statements (R2) and have positive (R1 - R2) difference scores. Conversely, bomb participants were expected to react more strongly to bomb statements (R2) and produce negative (R1 - R2) difference scores. The other variable was the difference between the combined mean reaction to drug (R1) and bomb (R2) statements and the reaction to neutral statements (i.e., ((R1 + R2)/2) – N). Guilty participants were expected to show greater differences between crime-related and neutral statements than

innocent participants.

Independent Variables

Two group membership indicator variables were created to distinguish among the three groups. One indicator variable (Guilty-Innocent) differentiated between the guilty (coded +1) and innocent participants (0). The second indicator variable (Drug-Bomb) distinguished between the Drug (+1) and Bomb groups (-1); innocent participants were coded 0.

The correlations between of the two group indicator variables and physiological indices of differential reactivity are presented in Table 4. The correlations indicate the extent to which the outcome measure discriminated between the groups and may be viewed as an index of predictive validity. For example, on the Guilty-Innocent indicator variable, guilty participants were coded +1 and innocent participants were coded 0. Thus, a positive correlation with this indicator variable indicated that the guilty group had a higher mean score on the outcome measure than the innocent group. Conversely, a negative correlation with this indicator variable indicated that the guilty group had a lower mean score on the outcome measure than the innocent group. We expected the ((R1+R2)/2) – N contrast to discriminate between guilty and innocent participants and have significant correlations with Guilt. We also expected the (R1-R2) contrast to discriminate between the drug and bomb groups and have significant correlations with Crime.

The correlations between discrete physiological features and group membership contrasts are consistent with the physiological reactions presented in Figures 1-8. The features most diagnostic of group membership were SC amplitude and SC area under the curve. These were followed by SBP, DBP, FPA, RLL, and PD. For example, SC peak amplitude discriminated between the two guilty groups (r = .633) and between the innocent group and the two guilty groups combined (r = .412). For all measures, the differences between the two guilty groups

Table 4. *Correlations between physiological features and group membership indicator variables (N = 354)*

Channel	Feature	Contrast	Guilt[a]	Crime[b]
Respiration	Line length 0-10s	((R1+R2)/2) - N	**.132**	**.107**
		(R1 – R2)	.023	**.294**
	Line length 6-12s	((R1+R2)/2) - N	.093	.036
		(R1 – R2)	.054	**.288**
Skin conductance	Peak amplitude	((R1+R2)/2) - N	**.412**	.037
		(R1 – R2)	.026	**.633**
	Area under the curve	((R1+R2)/2) - N	**.434**	.033
		(R1 – R2)	.037	**.656**
Finger pulse amplitude	Area under the curve	((R1+R2)/2) - N	**.217**	-.038
		(R1 – R2)	-.023	**.447**
	Duration	((R1+R2)/2) - N	**.234**	.022
		(R1 – R2)	-.026	**.313**
Systolic blood pressure	Peak amplitude	((R1+R2)/2) - N	**.291**	-.069
		(R1 – R2)	.010	**.536**
	Area under the curve	((R1+R2)/2) - N	**.318**	-.077
		(R1 – R2)	.023	**.571**
Diastolic blood pressure	Peak amplitude	((R1+R2)/2) - N	**.254**	-.055
		(R1 – R2)	-.009	**.469**
	Area under the curve	((R1+R2)/2) - N	**.292**	-.051
		(R1 – R2)	.043	**.475**
Heart rate	Maximum increase 0-8s	((R1+R2)/2) - N	-.086	.066
		(R1 – R2)	-.022	-.029
	Maximum decrease 0-16s	((R1+R2)/2) - N	**.171**	-.044
		(R1 – R2)	-.031	**.205**
Pupil diameter	Peak amplitude	((R1+R2)/2) - N	-.083	.012
		(R1 – R2)	-.009	.033
	Area under the curve	((R1+R2)/2) - N	-.050	.036
		(R1 – R2)	.035	**.139**

Note: For $N = 354$, $|r| > .107$ was significant at $p < .05$. Significant correlations are bolded.

[a]Guilt was a dichotomous variable that distinguished between innocent (coded 0) and guilty participants (coded 1).

[b]Crime was a trichotomous variable that distinguished among bomb (coded -1), innocent (coded 0), and drug (coded 1) groups.

(correlations with Crime) were greater than the differences between the innocent and guilty groups (correlations with Guilt). This was expected because the two guilty groups were on opposite sides of the Crime continuum (-1, 0, 1), and innocent group was sandwiched between them. The RCT was designed to produce that specific pattern of differences among groups.

Another objective of the present study was to evaluate less invasive alternatives to the cardiograph and skin conductance. Because cardiograph recordings were not obtained in the present study, it was not possible to test whether any of the alternative cardiovascular measures (BP, HR, LVET, SV, TPR) accounted for variance in group membership explained by the cardiograph. However, it was possible to determine if changes in pupil could be used in place of skin conductance. Whereas recordings of skin conductance require contact sensors, recordings of pupil size may be obtained unobtrusively with modern remote eye trackers.

To determine if skin conductance accounted for variance in group membership not already explained by changes in pupil size, I conducted multiple regression analyses that included both pupil and skin conductance measures. In one regression analysis, Guilt (innocent coded 0 and guilty coded 1) served as the dependent variable and the (R1+R2)/2 – N contrasts for pupil diameter area under the response curve and skin conductance area under the curve served as independent variables. The regression coefficient for skin conductance area under the curve was significant, B = .439, $p < .001$, which indicated that skin conductance accounted for variance in Guilt that was not explained by changes in pupil diameter. The regression coefficient for the pupil measures was not significant.

Another regression analysis was conducted where Crime (bomb coded -1, innocent coded 0, and drug coded 1) served as the dependent variable and the (R1-R2) contrasts for pupil diameter area under the curve and skin conductance area under the curve served as independent variables. Again, the regression coefficient for skin conductance was significant, B = .655, $p < .001$, whereas the coefficient for pupil diameter was not. These analyses were

consistent with the bivariate analyses reported in Table 3; they indicate that these pupil measures could not be substituted for skin conductance measures without a significant reduction in predictive validity.

Classification Analyses

Twenty-two of the variables listed in Table 3 were significantly correlated with one or both of the group membership indicator variables (Guilt or Crime). Since predictor variables are often highly intercorrelated, they provide partially redundant information about group membership. Rarely are more than 3 to 6 variables needed to capture all of the reliable diagnostic variance in a large set of potential predictor variables. To identify subsets of variables that predicted group membership, two separate all-possible-subset regressions were performed using the LEAPS package (Lumley, 2009) of the statistical program R (version 2.15.0; R Development Team, 2012). Multiple regression was used to identify diagnostic subsets of physiological measures for the statistical classifiers because multiple regression is mathematically equivalent to discriminant analysis when there are only two groups, and I was unaware of any statistical package that computes all-possible-subsets for discriminant analysis. I assumed that a variable considered diagnostic of group membership with regression statistics also would potentially be considered diagnostic in bagging and boosting classification models. One all-possible-subsets regression analysis was conducted with the significant ((R1+R2)/2) – N variables to discriminate between guilty and innocent participants (Guilt). The other was conducted with the R1-R2 variables to discriminate between the two guilty groups (Crime). Only participants in the Phase 1 sample were included in these analyses.

Each all-possible-subsets regression analysis produced a listing of the top five regression models with one predictor variable in the regression equation, two predictor variables in the equation, three predictors, and so forth. The analysis ended with the top five models that

contained eight predictor variables. There were large increments in R^2 between models with up to five variables, and little or no increase in R^2 thereafter. In addition, regression models with the same five physiological measures were among the top five listed subsets for predicting Guilt from ((R1+R2)/2) – N differences and predicting Crime from (R1-R2) differences. Those models used SC area under the curve, SBP area under the curve, DBP area under the curve, FPA duration, and RLL. Therefore, only the ((R1+R2)/2) – N and (R1-R2) difference scores for those five physiological measures were retained for possible inclusion in the decision models.

Standardization and Validation Samples

Bagging and boosting are designed for use on large data sets, and the largest feasible data set is recommended for development of the decision algorithms. To maximize the sample size for development of the decision model and its reliability, I restrained only as many cases as necessary to obtain reasonable estimates of accuracy on cross-validation (30 in each group) and used all of the remaining 254 cases as the standardization sample. The resulting standardization sample contained 89 drug, 91 bomb, and 94 innocent participants.

To provide a baseline model from which to evaluate ensemble classification methods bagging and boosting, discriminant analyses were run using standardization sample. Discriminant analysis yielded two functions. The first discriminant function accounted for more of the variance among the three treatment groups (R = .704) than did the second (R = .481). The proportion of variance in group membership explained by both functions was .697. The standardized discriminant function coefficients for the two functions are reported in Table 5.

The resulting classification of participants into groups is provided in Table 6. The percentages shown within parentheses are percentages of row totals. In the standardization sample, 69.7% of the 254 participants were classified correctly.

Table 5. *Standardized discriminant function coefficients for functions 1 and 2 in the ensemble standardization sample*

Channel	**Feature**	**Contrast**	**Function 1**	**Function 2**
Skin conductance	Area under the curve	(R1 – R2)	**.843**	.019
Systolic blood pressure	Area under the curve	(R1 – R2)	**.683**	.043
Diastolic blood pressure	Area under the curve	(R1 – R2)	**.555**	.058
Finger pulse amplitude	Duration	(R1 – R2)	**.458**	-.037
Respiration	Line length 0-10s	(R1 – R2)	**.294**	-.003
Skin conductance	Area under the curve	((R1+R2)/2) - N	.037	**.927**
Systolic blood pressure	Area under the curve	((R1+R2)/2) - N	-.117	**.690**
Diastolic blood pressure	Area under the curve	((R1+R2)/2) - N	-.090	**.588**
Finger pulse amplitude	Duration	((R1+R2)/2) - N	.056	**.384**

Note: Significant coefficients are bolded.

Table 6. *Classifications of participants in the standardization sample*

		Decision		
Condition	N	"Innocent"	"Drug"	"Bomb"
Innocent	94	65 (69.1%)	17 (18.1%)	12 (12.8%)
Drug	89	23 (25.8%)	61 (68.5%)	5 (5.6%)
Bomb	81	18 (22.2%)	5 (6.2%)	58 (71.6%)

Validation Sample

The discriminant functions from the standardization sample were used to compute the probabilities of group membership for each participant in the validation sample. The results for the validation sample are presented in Table 7. On cross-validation, overall accuracy increased from 69.7% to 76.7% (+7.0%). Accuracy decreased on innocent (-2.4%) but increased on drug participants (+11.5%) and bomb participants (+11.7%). To find that accuracy was higher on cross-validation than in the standardization sample was not expected.

Ensemble Classification

The (R1-R2) and (R1+R2)/2 – N contrasts for respiration, SC, SBP, DBP, FPA features were made available for inclusion in the discriminant functions. To compare the accuracy of the

Table 7. *Classifications of participants in the validation sample*

		Decision		
Condition	N		"Drug"	"Bomb"
Innocent	30	20 (66.7%)	6 (20.0%)	4 (13.3%)
Drug	30	4 (13.3%)	24 (80.0%)	2 (6.7%)
Bomb	30	5 (16.7%)	0 (0.0%)	25 (83.3%)

ensemble classifiers to that achieved with discriminant analysis, the same contrasts were used for bagging and boosting. Several attempts were made to improve each of the bagging and boosting algorithms. For initial attempts, all nine difference scores selected by multiple regression were available to the bagging and boosting algorithms. In the second iteration, DBP measures were constrained in an attempt to improve predictive strength on cross-validation. FPA measures were removed from the pool of potential predictors in the third iteration. A fourth iteration using variables identified by principle components analysis was performed with the boosting algorithm. The percentages of cases correctly classified by bagging and boosting algorithms for all iterations are presented with results of discriminant function analysis for comparison in Table 8.

Bagging algorithms yielded generally poor results, with 30.0% of cases correctly assigned in the Bagging 1 standardization sample and 61.1% correctly assigned in the validation sample.

Such a large improvement in the classification of the validation sample was unexpected. Although the Bagging 2 and Bagging 3 models yielded slightly better classification rates in standardization sample (34.4% and 34.7%, respectively), cases in the validation sample were classified accurately only 29.7% of the time. Bagging misclassified innocent subjects more than other groups, especially in the validation sample, although it correctly classified at least 90% of those in the drug group. Bagging misclassified bomb participants more than other groups in standardization and validation samples. The bagging algorithm's decline in classification performance when the accessible classifiers were constrained was not expected. Bagging

Table 8. *Percent of cases classified correctly for discriminant function analysis (DFA) and ensemble methods*

		Standardization Sample (*N* = 264)				Validation Sample (*N* = 90)				
Method	Model	Innocent	Drug	Bomb	Mean	Innocent	Drug	Bomb	Mean	Mean Change
DFA	0	69.1	68.5	71.6	69.7	66.7	80.0	83.3	76.7	+7.0
Bagging	1	25.5	56.2	8.6	30.0	13.3	93.3	76.7	61.1	+31.1
Bagging	2	21.3	71.9	9.9	34.4	3.3	90.0	76.7	29.7	-4.7
Bagging	3	21.3	73.0	9.9	34.7	3.3	90.0	76.7	29.7	-4.7
Boosting	1	88.3	91.0	85.2	88.2	53.4	80.0	80.0	71.1	-17.1
Boosting	2	91.5	88.8	88.9	89.7	60.0	80.0	73.3	71.1	-18.6
Boosting	3	80.0	87.7	87.7	85.1	53.3	73.3	63.3	63.3	-21.8
Boosting	4	83.0	82.0	90.1	85.0	53.3	63.3	60.0	58.9	-26.2

Notes

Model 1: Variables included: (R1-R2) and [(R1+R2)/2 - N] contrasts for SC, SBP, DBP, FPA, and the (R1-R2) contrast for respiration

Model 2: Variables included: (R1-R2) and [(R1+R2)/2 - N] contrasts for SC, SBP, FPA, and the (R1-R2) contrast for respiration

Model 3: Variables included: (R1-R2) and [(R1+R2)/2 - N] contrasts for SC, SBP, and the (R1-R2) contrast for respiration

Model 4: Variables included principal components derived from the (R1-R2) and [(R1+R2)/2 - N] contrasts for SC, SBP, DBP, FPA, and the (R1-R2) contrast for respiration

performed poorly in comparison to DFA, yielding a 39.7% decrease in performance for the standardization sample and a 15.5% decrease in the validation sample. Bagging 1, the best bagging model, assigned 100% of the weight to the (R1-R2) contrast for skin conductance.

As noted above, Boosting models 1-3 were provided the same subsets of physiological measures as Bagging models 1-3. Boosting models generally outperformed the bagging classifiers. The best boosting algorithm, produced when diastolic blood pressure variables were omitted from the available pool of predictors (Boosting 2), correctly classified 89.7% of the standardization sample (+20.0% from DFA) and 71.1% of the validation sample (-5.6% from DFA). Boosting 2 assigned 91.9% of the weight to the (R1-R2) contrast score and 9.1% to the (R1-R2) contrast for systolic blood pressure. The Boosting 1 algorithm correctly classified 88.2% and 71.1% of subjects in standardization and validation samples, respectively. Boosting 3 performed worse on standardization (85.1%) and validation (63.3%) samples. In general, accuracy declined substantially on cross-validation, particularly with Innocent subjects.

As noted above, Boosting 4 used classifiers identified by Principle Component Analysis (PCA). The intent was to reduce the dependency among predictor variables by replacing the original physiological contrasts with weighted sums (components) that were orthogonalized by Varimax rotation (Gorsuch, 1970). The analysis identified four principal components . The first component consisted of R1-R2 contrasts for SC, SBP, DBP, and FPA . The second component consisted of (R1+R2)/2 – N contrasts for SC, SBP, DBP, and FPA. The third component consisted of mean of the SC, SBP, DBP, and FPA absolute differences between R1 and R2, and the fourth component consisted of the R1-R2 contrast for RLL used in all previous decision models. Boosting 4 produced 85.0% and 58.9% mean accuracy for the standardization and validation samples, respectively.

Effects of Trimming Branches from Boosting Decision Trees

In boosting, branches are iterations or rounds of the statistical method's attempts to fit the data. By default, the boosting algorithm uses 10 branches (decision models) to classify cases into groups. To determine if fewer branches would improve the generalizability of the boosting method, we explored the effects of trimming the algorithm from 10 to 5 or 2 branches. When the algorithm was trimmed to 5 branches, percentages of correctly assigned cases decreased slightly on standardization and validation samples to 82.1% and 63.2%, respectively. When M = 2, accuracy again declined , decreasing to 68.6% for standardization and 68.9% for validation samples. When M = 5 and SC and DBP were omitted, the algorithm produced accuracy rates of 78.4% and 70.0% for standardization and validation samples, respectively. These rates were comparable to DFA and show less shrinkage than in other boosting attempts.

Because discriminant analysis capitalizes on chance to maximize separation of the groups in the standardization sample, there is good reason to expect its performance to deteriorate on cross-validation. This phenomenon is measured as "shrinkage" (McNemar, 1969). In the present study, the accuracy of discriminant analysis was higher in the validation sample than the standardization sample. This suggests that sampling error was responsible; by chance, the standardization sample contained more marginal cases than did the validation sample. We rerandomized the split between standardization and validation samples and observed the expected higher accuracy in the standardization than the validation sample, but the relative differences in accuracy among classification methods (discriminant analysis, bagging, and boosting) and models remained. Therefore, only the results with the original subdivision into standardization and validation samples are reported above.

CHAPTER 4

DISCUSSION

Relevant Comparison Test

We conducted a mock crime experiment to evaluate an automated RCT for possible use at ports of entry. Guilty participants transported either a substance that appeared to be illegal drugs or a device that appeared to be a bomb. Innocent participants committed neither crime.

The RCT was based on the assumption that Drug participants would react more strongly to questions about the drugs, Bomb participants would react more strongly to questions about the bomb, and innocent participants would show little difference in their reactions to drug and bomb questions. These predictions were confirmed for respiration, SC, FPA, SBP, DBP, HR, PD, and, to a lesser extent, TPR. Deception was associated with suppressed respirations, increased SC, decreased FPA (vasoconstriction), increased SBP, increased DBP, elevated tonic HR levels, phasic cardiac deceleration, and pupil enlargement. The nature of the effects on those measures is consistent with a large literature on the CQT (Kircher & Raskin, 2001; Podlesny & Raskin, 1977; Raskin & Hare, 1978; Raskin & Kircher, 2014). The findings are consistent with the idea that the observed effects are predominantly, though not exclusively, a noradrenergic reaction of the sympathetic nervous system.

In contrast to the CQT, the RCT contained no probable-lie or directed-lie comparison questions obviating the need to provide the participant with a rationale for asking such questions. There was no concern that the innocent participant would fail to accept that rationale, fail to react appropriately to the comparison questions, and fail the test. The lack of

comparison questions simplified the pretest, and reduced the time to describe the protocol and present instructions to the participant. The pretest portion of the RCT took less than 4 min.

The RCT also addresses concerns expressed by critics of probable-lie and directed-lie tests that the comparison questions do not provide a proper control condition for evaluating reactions to relevant questions (Ben-Shakhar & Furedy, 1990; Lykken, 1998). Each relevant issue on the RCT provided a standard against which reactions to the other relevant issue could be compared. For a bomb participant, reactions to questions concerning the drugs provided an indication of how that individual would have responded to accusatory questions about the bomb if the bomb questions had been answered truthfully. The converse was true for participants guilty of transporting drugs. For innocent participants, there should have been no perceived difference in the importance of the two sets of relevant questions and no difference in the relative strength of reactions to the two sets of questions.

Proper scientific controls are designed to assure the internal validity of experiments. However, a polygraph test is not an experiment. It is a psychological evaluation designed to determine if a person is telling the truth. At issue is the criterion-related validity of the test; that is, does it detect deception? Although it is unnecessary to address this particular concern of critics, the RCT does so, and if the RCT were to work as well or better than a CQT, there would be some value in using it to defuse this particular issue and gain broader support in the scientific community for polygraph testing.

RCT addresses some concerns expressed by critics of comparison-question polygraph techniques, but the accuracy of the RCT was about 20% lower than the accuracy typically obtained with the CQT in laboratory experiments. Whereas the RCT classified participants into three categories, prior research on the CQT classified participants into only two categories. The chance probability of a correct decision by the RCT was approximately 33%, whereas chance in research on comparison-question tests is typically 50%. The gain in accuracy over chance

achieved by the RCT in the present study was 35% to 40%, which is comparable to that of comparison-question tests.

The RCT data for Phase 1 and Phase 2 accounted for 63% of the variance in group membership. In one recent study, 63% of the variance in group membership was explained by the PLT, and 45% of the variance was explained by the DLT (Bell, Kircher, & Bernhardt, 2008). In a lens model analysis of polygraphs from confirmed criminal cases, 36% of the variance in group membership was explained by the optimal combination of physiological measures (Kircher, Kristjansson, Gardner, & Webb, 2012). The proportion of variance in group membership in the present study was as least as large as that achieved in other studies of other polygraph techniques. Together, these results indicate that the mock crime procedures in the present study produced large diagnostic effects on the physiological measures that were comparable to those obtained in prior research on the CQT. Despite these large effects, the accuracy rates were not high enough to recommend use of the RCT for screening at ports of entry without further development.

Automation

The present study used an automated test protocol that produced large effects on physiological measures used by polygraph examiners to detect deception. In an experiment by Honts and Amato (2007), an interviewer first met with the examinee and discussed the matter under investigation. At that point, a tape recorded human voice presented the test questions to the participant. In the present study, a research assistant attached the sensors to the participant, calibrated the Finometer, and monitored signal quality, but the pretest instructions and in-test presentation of test items were completely automated. In Phase 1 of the present study, a prerecorded synthetic voice conducted the examination. In Phase 2, a prerecorded human voice conducted the examination. Although other procedural differences between

Phase 1 and Phase 2 were confounded with the synthetic versus human voice manipulation, substantive differences were observed between the two phases only in the quality of the respiration recordings, and those effects probably were due to changes in participants' posture. Given the large sample sizes in Phase 1 and Phase 2, the present findings suggest that effects of voice on physiological reactions to test questions are likely to be small and inconsequential.

The present findings indicated that certain types of polygraph tests can be fully automated, especially those used in screening applications (Hont & Amato, 2007). The polygraph examiner is a major source of uncontrolled variance that may affect the reliability of the test (Ben-Shakhar & Furedy, 1990; Lykken, 1998; National Research Council, 2003). Automation of the pretest and in-test phases of the polygraph test would address that concern. On the other hand, automation virtually eliminates the social competition between the examiner and examinee that characterizes the traditional CQT. If competition between the examinee and examiner is an essential ingredient of the deceptive context, then some level of interaction with an examiner could improve the accuracy of the RCT (Podlesny & Raskin, 1977).

Less Invasive Alternatives to Skin Conductance and the Cardiograph

Effects of deception on SC, respiration, and FPA during an RCT were similar to those obtained in prior research on CQT and DLT. The present study also investigated physiological measures that are not recorded in traditional polygraph tests. These include pupil size and several new cardiovascular measures.

Pupil Size

The observed effects of deception on changes in pupil size were considerably smaller in the present study (3.5% of the variance) than in a previous study of the PLT (10.0% of the variance)(Webb et al., 2009). Equipment problems contributed to poor effect size in the present

study. Signal quality was poor from the HawkEye (Acunetx, Torrence, CA) in Phase 1, partially because the equipment design required that participants recline in a chair in a low-lit room, causing some participants to become drowsy and close their eyelids. The HawkEye was replaced by an Arrington EyeFrame tracker in the Phase 2, allowing participants to sit upright and minimize drowsiness. Despite efforts to improve the quality of the pupil size signal in Phase 2, a systematic visual inspection of all pupil recordings revealed numerous artifacts and signal losses. Ratings were made of pupil signal quality and used to determine if better quality signals yielded better predictors of group membership. Those efforts improved the correlations with group membership but not appreciably. We are uncertain whether the small effects on pupil size indicate that it is not useful for the present application of the RCT, or the instrumentation in both phases was inadequate, and/or techniques used by lab personnel to position the camera and light source and record changes in pupil size were deficient. Considering prior research (e.g., Bradley & Janisse, 1981; Webb et al., 2009) and the ambiguity of the present findings, future efforts to develop noncontact sensors for deception detection should include measures of pupil dilation.

Cardiovascular Measures

The present study used a Finometer Pro to record SBP, DBP, SV, LVET, and TPR. The systolic points of the arterial pressure curve also were used to measure HR. The results from the Finometer were mixed. SBP and DBP were highly correlated with group membership and contributed substantially to the discriminant functions. The present findings for SBP and DBP are consistent with results obtained with the Finapres (Bell, Kircher, & Bernhardt, 2008; Craig, Raskin, & Kircher, 2011; Podlesny & Kircher, 1999). The older Finapres and newer Finometer used the same vascular unloading principle to monitor changes in arterial pressure continuously from a low-pressure finger cuff. Although the Finometer was expensive ($32,000), use of less

invasive technology would be preferable to the cardiograph in a port-of-entry screening context. The SV and LVET measures obtained from the Finometer were not sensitive to the deception manipulations. Examination of second-by-second measures of TPR revealed that TPR decreased more when the participant was deceptive than truthful. The effect on TPR began about 8 s after question onset and lasted for at least 6 s. At the same time, HR was decelerating and SBP and DBP were at their highest levels. The temporal relationship among these measures suggests that the decrease in TPR and cardiac deceleration may have been the body's attempt to control the rapid increase in blood pressure that began one or two seconds earlier. Others have observed decreases in TPR during real fear (Stemmler, Heldmann, Pauls, & Scherer, 2001) and imagined fear (Sinha, Lovallo, & Parsons, 1992; Stemmler et al., 2001) but not during anger inductions. Stemmler et al. suggested that the decrease in TPR is a component of a defensive reflex to prepare the individual for flight.

Ensemble Classification Methods

Ensemble classification methods were tested as an alternative to discriminant analysis. Boosting, bagging, and discriminant analysis weighted SC variables over other available variables in all decision models. Boosting classified participants with higher rates of accuracy than did bagging, but neither ensemble classification method outperformed DFA. As compared to discriminant analysis, the best boosting algorithm classified more of the standardization sample correctly, but it performed worse on cross-validation. Bagging algorithms performed poorly. DFA is the preferred decision model because it performed the best on cross-validation.

As noted above, several findings were counterintuitive. First, bagging and boosting performed less impressively than expected. There are several possible explanations for this finding. Bagging and boosting were designed for use on large data sets. Although we collected a large amount of data, focusing on 9 measures of physiological change, the data set may not

have been large enough for the ensemble methods. More specifically, ensemble methods require a fair amount of unexplained variance and weak classifiers for the algorithms to be effective. Because we had already identified classifiers that were known to be effective, it is likely that the algorithms did not have sufficient unexplained variance in which to work. Essentially, by identifying and using predictors that were highly correlated with group membership , we did the work that the ensemble methods were designed to accomplish, negating any impact of the algorithms. This also explains why our attempts to constrain the available classifiers as well as orthogonalize the predictors yielded poorer results. The set of four orthogonalized variables yielded the worst results when used by the boosting algorithm because they contained less error than did the original physiological measures used in other statistical classifiers.

We resplit the sample to determine if our validation sample was somehow biased in comparison with the standardization sample. Predictably, for DFA, those attempts increased accuracy in the standardization sample and reduced accuracy in the validation sample, but these attempts yielded little difference in results for bagging and boosting models. We also experimented with the boosting algorithm to determine what might improve the predictive ability of the algorithm in future attempts. First, we removed the strongest predictors, SC and SBP, to allow the algorithm access to only weaker predictors. The algorithm produced nearly the same results in the standardization and validation samples as when those classifiers were included, so removing the strongest classifiers did not appear to be effective. Second, however, it is interesting that the algorithm did just as well in classifying the data when it did not have access to the SC and SBP measures, suggesting that the algorithm does benefit from access to weaker predictors.

As noted above, we observed considerable shrinkage in bagging and boosting attempts. Validation sample accuracy rates fell considerably lower than in standardization samples for

both bagging and boosting, while DFA's accuracy increased slightly in the validation sample. This result suggests that bagging and boosting may overfit the data and produce a less generalizable algorithm. However, although we attempted to reserve a number of cases in each group that we believed would maximize the size of the standardization sample but still allow a sufficient number of cases to test the algorithms, it may be that the standardization sample still was too small to represent populations of drug, bomb, and innocent participants.

Limitations

A potential problem with the RCT is that it should fail if the relevant issues are equally salient and the participant is deceptive to both sets of relevant questions. In that case, the participant should react strongly to both relevant questions. There would be no difference between reactions to two relevant questions, and the participant would be considered truthful. The RCT would be most appropriate in situations where it is unlikely that the participant would be deceptive to both relevant issues. However, the RCT might still be useful as a law enforcement preemployment polygraph test (LEPET) if one relevant issue were a serious crime with an extremely low base rate of occurrence such as espionage, and the other issue was a crime with a relatively high base rate of occurrence such as illegal drug use. The innocent participant should react similarly to the two sets of relevant questions, since strong reactions to either question would have undesirable consequences. A person who lies about one or the other relevant issue would be expected to react more strongly to that relevant question and fail the test on that issue. Conversely, a person who is deceptive to both sets of questions probably would react more strongly to questions concerning the more serious offense and fail the test on that issue.

Countermeasures are another potential concern for the RCT. The CQT protects against drug countermeasures that generally reduce (or enhance) reactivity because a lack of difference

between reactions to relevant and comparison questions on the CQT produces an inconclusive outcome. In contrast, a lack of difference between reactions to the two relevant questions on an RCT results in a truthful outcome. Although there is little evidence that pharmacological countermeasures can be used to defeat comparison-question tests (Honts , Raskin, & Kircher, 2002), they may be effective against an RCT. Research is needed to address this concern.

The present study did not obtain measures of cardiovascular activity from the cardiograph because it would have limited the number of test questions we presented and lengthened the test. Another limitation was that we could not test if the BP measures accounted for variance in group membership that would have been explained by the cardiograph. Previous research with the probable-lie comparison test has indicated that measures of BP obtained with the vascular-unloading finger cuff are more diagnostic than measures obtained with the cardiograph (Podlesny & Kircher, 1999). Additional research is needed to determine if similar findings are obtained with the RCT.

There were limitations related to ensemble classification methods in the present study. While the sample size was large and produced a large amount of physiological data, the data were limited to nine measures that were found to have good predictive ability in previous research on polygraph techniques. Ensemble methods require large data sets and access to weak predictors, and they generally are used in exploratory studies when little is known about the predictors and their relationship to group membership. A better use of ensemble methods might be to provide the algorithms access to the entire set of physiological outputs, most of which are only weakly related to group membership. In that case, the algorithms could identify classifiers that could be useful in building the final classification algorithm. It may be that the ensemble methods would still underperform in this circumstance, however; the data set may need to be exponentially larger than in the present study in order for the ensemble methods to improve over DFA. The size of the validation sample also was problematic. Cross-validation

procedures available in R for bagging and boosting might be a more effective alternative than the methods used in the present study because a hold-out sample is not required. Finally, bagging and boosting may not be the best ensemble methods to use with physiological measures of arousal. More research is required to fully address this issue.

General Conclusions

The present study evaluated a brief, automated psychophysiological test for deception for use at ports of entry. The study introduced a new test format, the RCT. The RCT has several conceptual and practical advantages over traditional comparison-question formats. The RCT requires little psychological preparation of the examinee prior to the test, lends itself to automation, and addresses certain criticisms of comparison-question formats. Although the accuracy achieved by the RCT was lower than that typically obtained in mock crime studies of the CQT, the magnitude of effects on physiological measures was just as high. The latter findings suggest that the RCT merits additional research. With high quality recordings of pupil size or other sources of diagnostic information and use of other question sets, the RCT may ultimately play a role in some screening programs. The study also evaluated new statistical analyses for classification of subjects into groups. Bagging and boosting algorithms offered no advantage over DFA in the present study. Further investigation is required to fully assess the potential benefit of ensemble methods for classifying subjects in this area.

APPENDIX A

COMPUTERIZED SCREENING SYSTEM

INCLUSION/EXCLUSION CRITERIA

Questions to Be Asked of Participant

		Yes	No
1.	Are you feeling sick today?	___	___
2.	Have you had at least six hours sleep last night?	___	___
3.	Are you at least 18 years of age?	___	___
4.	Are you taking prescription medications for heart or psychiatric conditions?	___	___

Questions to Be Filled Out by the Research Assistant

		Yes	No
5.	Does the participant appear to be intoxicated or under the influence of drugs?	___	___
6.	Does the participant appear to be overly tired?	___	___
7.	Does the participant appear to be ill?	___	___

To be eligible for participation, the answers to all questions except question #2 and #3 must be "No." The answer to question #2 and #3 must be "Yes."

If the participant fails the inclusion/exclusion criteria screening, please thank the participant for coming, and explain they are not eligible to be in the study. Pay the subject $15.

APPENDIX B

AUTOMATED POLYGRAPH SCRIPT

"You are about to be asked a series of questions. The questions will be repeated several times. The series will include questions about transporting illegal drugs and bomb parts. The test will take about 15 minutes. After you complete the test, the computer will analyze your data and indicate if you passed the test."

"This is how the polygraph works. If you were walking alone down a dark street late at night and heard a noise you didn't expect, you'd stop. You'd listen. Your heart would begin to pump harder, your hands would begin to sweat, and your blood pressure would go up. This is called the fight-or-flight response. Your body is preparing you to escape from the situation or engage in a fight. All of these physiological changes occur automatically whenever you perceive a threat to your well-being. You don't have to think about it, it just happens. Your body automatically gets you ready to deal with the threat."

"Now, if we ask you a question about something important and you lie, you will be concerned that we will find out. It is natural for a person to be concerned, or even fearful, that their lies will be detected. If you intend to lie to a question, that question will pose a threat to you, just like a noise in the dark you didn't expect. It is something to be feared or concerned about."

"During the polygraph test, we will record your heart rate, blood pressure, sweating, breathing, and pupil size. If you lie to a question, you will be concerned that your lie will be detected, and we would expect to see the fight-or-flight response. On the other hand, if you are completely truthful, there is nothing to be concerned about because there is no lie to detect. If you are completely truthful to all the questions, we would not expect to see the fight-or-flight response."

"Do you understand?"

[If participant answers, "No," the experimenter asks if the participant wants the information repeated, or if the participant has questions.]

"In a minute or so, you will feel the cuff on your finger inflate. From that point onward until the test is over, it is important that you continue to breathe normally and remain as still as possible. Answer each question clearly either "Yes" or "No." Avoid moving your head, hands, or feet. If you move too much, the computer will caution you to sit quietly. If you move too much, the computer may add questions to the series, and that will lengthen the test. If you sit quietly throughout the test, it will take about 15 minutes."

"Do you have any questions?" "Are you ready to begin?" [after participant "Yes,"] "Please sit quietly, and answer each question "Yes" or "No."

REFERENCES

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Examiner*, *15*(1), 6-11.

Bell, B.G., Kircher, J.C., & Bernhardt, P.C. (2008). New measures improve the accuracy of the directed-lie test when detecting deception using a mock crime. *Physiology and Behavior, 94,* 331-340.

Ben-Shakhar, G., & Furedy, J.J. (1990). Theories and applications in the detection of deception: A psychophysiological and international perspective. New York: Springer-Verlag.

Bradly, M. T. & Janisse, M. P. (1981). Accuracy Demonstrations, Threat, and the Detection of Deception: Cardiovascular, Electrodermal, and Pupillary Measures. *Psychophysiology, 18*(3), 307–315.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Brownlie, C., Johnson, G. J., & Knill, B. (2002). Validation Study of the Relevant/Irrelevant Screening Format. Draft report to the Department of Defense Polygraph Institute.

Bühlmann, P., Yu, B. (2002). Analyzing bagging. *Annals of Statistics, 30,* 927-961.

Craig, R. A., Raskin, D. C., & Kircher, J. C. (2011). The use of physiological measures to detect deception in juveniles. *Polygraph, 40*(2).

Dollins, A. B., Krapohl, D. J., & Dutton, D. W. (2000). A comparison of computer programs designed to evaluate psychophysiological detection of deception examinations. *Polygraph, 29*(3), 237-257.

Freund, Y., & Schapire, R. (1996). Experiment with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning* (pp. 148–156). San Francisco: Morgan Kaufmann.

Gorsuch, R. L. (1970). A comparison of biquartimin, maxplane, promax, and varimax. *Educational and Psychological Measurement, Vol. 30*(4), 861-872.

Honts, C. R., & Amato, S. (2007). Automation of a screening polygraph test increases accuracy. *Psychology, Crime & Law, 13*(2), 187-199.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (2002). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. Kaye, M. J. Saks, & J.

Saunders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (2nd Edition, pp. 446-483). St. Paul, MN: West Publishing.

Horowitz, S. W., Kircher, J. C., Honts, C. R., & Raskin, D. C. (1997). The role of control questions in physiological detection of deception. *Psychophysiology, 34,* 108-115.

Kircher, J. C., Kristjansson, S., Gardner, M. K., & Webb, A., K. (2012). Human and computer decision making in the psychophysiological detection of deception. *Polygraph, 41*(2), 77-126.

Kircher, J. C., & Raskin, D. C. (1988a). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology, 73,* 291-302.

Kircher, J. C., & Raskin, D. C. (1988b). Comment: Baserates and the statistical precision of polygraph tests in various applications. *Statistical Science, 2*, 226-228.

Kircher, J. C. & *Raskin*, D. C. (*2001*). Computer methods for the psychophysiological detection of deception. In M. Kleiner (Ed.), *Handbook of Polygraph Testing* (pp. 287-326), London, England: Academic Press.

Kircher, J. C., Woltz, D. J., Bell, B. G., & Bernhardt, P. C. (1998). *Effects of audiovisual presentations of test questions during relevant-irrelevant polygraph examinations and new measures.* Final report to the U. S. Government. Salt Lake City: University of Utah, Department of Educational Psychology. Available from the first author.

Krapohl, D. J. (2001). The polygraph in personnel screening. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp 217-236). London: Academic Press.

Lumley, T. (2009). Using Fortran code by Alan Miller (2009). leaps: regression subset selection. R package version 2.9. http://CRAN.R-project.org/package=leaps.

Lykken, D. T. (1998). *A tremor in the blood: Uses and abuses of the lie detector.* Reading: Perseus.

McNemar, Q. (1969). *Psychological Statistics* (4th Ed.). New York, NY: Wiley.

National Commission on Terrorist Attacks upon the United States, Kean, T. H., & Hamilton, L. (2004). *The 9/11 Commission report: Final report of the National Commission on Terrorist Attacks upon the United States*. Washington, D.C.: National Commission on Terrorist Attacks upon the United States.

O'Sullivan, M., Frank, M. G., & Hurley, C. M. (2009). Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior, 33*(6), 530-538.

Podlesny, J. A., & Kircher, J. C. (1999). The Finapres (volume clamp) recording method in psychophysiological detection of deception examinations: Experimental comparison with the cardiograph method. *Forensic Science Communication, 1*(3), 1-17.

Podlesny, J. A. & Raskin, D. C. (1977). Physiological measures and the detection of

deception. *Psychological Bulletin, 84* (4), 782-799.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raskin, D. C. & Kircher, J. C. (2014). Validity of polygraph techniques and decision methods. In D. C. Raskin, C. R. Honts, & J. C. Kircher (Eds.), *Credibility assessment: Scientific research and applications.*

Raskin, D.C., Kircher, J.C., Honts, C.R., & Horowitz, S.W. (1988). *A study of the validity of polygraph examinations in criminal investigation.* Final report to the National Institute of Justice (Grant No. 85-IJ-CX-0040). Salt Lake City: University of Utah, Department of Psychology.

Raskin, D. C., & Hare, R. D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology, 15(2),* 126-136.

Seghetti, L. (2014). *Border security: Immigration inspections at ports of entry*. Washington, DC: Congressional Research Service.

Sinha, R., Lovallo, W. R., & Parsons, O.A. (1992). Cardiovascular differentiation of emotions. *Psychosomatic Medicine, 54*, 433-435.

Stemmler, G., Heldmann, M., Pauls, C. A., & Scherer, T. (2001). Constraints for emotion specificity in fear and anger: The context counts. *Psychophysiology, 38,* 275-291.

Webb, A. K, Honts, C. R., Kircher, J. C., Bernhardt, P.C., & Cook, A. E. (2009). Effectiveness of pupil diameter in a probable-lie comparison question test for deception. *Legal and Criminal Psychology, 14*(2), 279-292.